Rapport soumis aux rapporteurs, dans le but de sanctionner le
dossier pour l'obtention du grade de
Docteur en informatique
de
l'Université Paris-Nord

Acquisition automatique de schémas de sous-catégorisation à partir de corpus bruts

Cédric Messiant, le 5 septembre 2010

Mots-clés : verbe, sous-catégorisation, lexique, acquisition automatique, syntaxe, classes sémantiques

11h

11h45 Alexis Nasr
11h56 Núria Bel

12h09 Jacques Blanc-Tanon DGA
12h20 Antoine Rozelen-Knop ⎫
12h31 Addin Nazarenko ⎬ locaux
12h40 Thierry Poibeau ⎭
12h45 Eric

Table des matières

Table des matières i

Liste des figures v

Liste des tableaux vi

Liste des abréviations ix

1 Introduction 1
- 1.1 Domaine et motivations . . . 1
- 1.2 Enjeux . . . 4
- 1.3 Plan de la thèse . . . 4

2 Description du verbe pour l'élaboration de lexiques 7
- 2.1 La notion d'entrée lexicale . . . 8
- 2.2 Prédicat et structure argumentale . . . 9
- 2.3 Les différentes composantes des schémas prédicatifs . . . 10
  - 2.3.1 Exemple introductif . . . 10
  - 2.3.2 Distinction argument/modifieur . . . 11
  - 2.3.3 Schéma de sous-catégorisation . . . 12
  - 2.3.4 Grille thématique . . . 13
  - 2.3.5 Restrictions de sélection sur les arguments . . . 15
- 2.4 Alternances . . . 16
- 2.5 Les contours flous des schémas prédicatifs . . . 18
  - 2.5.1 Polysémie et entrées lexicales . . . 18
  - 2.5.2 Retour sur la distinction argument/modifieur . . . 18
  - 2.5.3 Formes pronominales . . . 19
  - 2.5.4 Verbes supports et expressions figées ou semi-figées . . . 21
- 2.6 Conclusion : vers l'acquisition automatique de schémas prédicatifs . . . 22

3 Les ressources existantes pour le français 23
- 3.1 Le Lexique-Grammaire . . . 24
- 3.2 Les ressources dérivées du Lexique-Grammaire : SynLex et LGLex . . . 28
  - 3.2.1 Synlex . . . 28
  - 3.2.2 LGLex . . . 29
- 3.3 DicoValence . . . 30
- 3.4 Le Lefff . . . 33
- 3.5 D'autres dictionnaires électroniques pour le français . . . 34
  - 3.5.1 Le dictionnaire syntaxique des Verbes Français . . . 34

3.5.2 LexValf . . . 36
3.5.3 Trésor de la Langue Française informatisé . . . 36
3.6 TreeLex, un lexique de sous-catégorisation acquis à partir d'un corpus arboré . . . 37
3.7 Conclusion . . . 39

4 Les méthodes d'acquisition automatique de schémas de sous-catégorisation 41
4.1 Méthodes d'évaluation de ressources lexicales . . . 43
4.1.1 Évaluation intrinsèque et évaluation extrinsèque . . . 43
4.1.2 Évaluation intrinsèque : évaluation quantitative . . . 44
4.1.3 Évaluation intrinsèque : évaluation qualitative . . . 45
4.1.4 Discussion . . . 45
4.2 Schéma global pour l'acquisition automatique de schémas de sous-catégorisation . . . 46
4.2.1 Étape 1 : identification des verbes en corpus . . . 46
4.2.2 Étape 2 : identification des compléments . . . 47
4.2.3 Étape 3 : association entre ces constructions locales et des schémas de sous-catégorisation . . . 47
4.2.4 Étape 4 : correction d'erreurs . . . 48
4.3 Les premiers travaux d'acquisition automatique de SSC . . 48
4.3.1 Jeter les bases de l'acquisition automatique de SSC à partir de corpus . . . 49
4.3.2 Mieux exploiter les corpus en les étiquetant *a priori* . . . 50
4.4 Acquisition de SSC à large échelle . . . 51
4.5 Travaux d'acquisition pour le français . . . 54
4.6 Conclusion . . . 55

5 Description du système d'acquisition de SSC pour le français 57
5.1 Architecture générale d'ASSCi . . . 58
5.2 Pré-traitements . . . 59
5.2.1 Lemmatisation et étiquetage morpho-syntaxique : TreeTagger 59
5.2.2 Analyse syntaxique de surface : Syntex . . . 62
5.3 Validation des verbes . . . 66
5.4 Description des modules d'ASSCi . . . 68
5.4.1 Extracteur de pré-schémas de sous-catégorisation . . . 68
5.4.2 Constructeur de schémas de sous-catégorisation candidats . . 74
5.4.3 Filtre des schémas non pertinents . . . 77
5.5 Conclusion . . . 82

6 Utilisation et évaluation des méthodes d'acquisition automatique de SSC 85
6.1 Acquisition d'un lexique de sous-catégorisation à large couverture pour le français . . . 86
6.1.1 Matériel et méthode . . . 86
6.1.2 Présentation du lexique . . . 88
6.1.3 Une interface web de consultation du lexique . . . 93
6.2 Évaluation de LexSchem . . . 95
6.2.1 Évaluation quantitative : comparaison avec d'autres ressources 95
6.2.2 Évaluation qualitative . . . 99

6.2.3 Conclusion sur l'évaluation . . . 104
6.3 Adaptabilité des méthodes d'acquisition automatique . . . 104
6.3.1 Acquisition de SSC de verbes pour une langue de spécialité . 104
6.3.2 Acquisition de SSC de prédicats non verbaux . . . 107
6.4 Des outils pour le paramétrage des lexiques . . . 108
6.4.1 Filtrage du lexique . . . 108
6.4.2 Formats disponibles pour le lexique . . . 109
6.4.3 Proposition de nouveaux schémas . . . 110
6.5 Conclusion . . . 111

7 Production de classes de verbes sur la base de leur comportement syntaxique 113
7.1 Problématique . . . 114
7.2 Travaux antérieurs . . . 115
7.3 Deux expériences pour l'acquisition de classes sémantiques de verbes . . . 116
7.3.1 Expérience 1 : classification par regroupement ascendant hiérarchique . . . 117
7.3.2 Expérience 2 : classification par regroupement spectral . . . 117
7.4 Discussion . . . 119
7.5 Conclusion . . . 120

8 Conclusion 121
8.1 Rappel des enjeux . . . 121
8.2 Contribution . . . 121
8.3 Perspectives . . . 123

A Liste des étiquettes de TreeTagger pour le français 127

B Exemple d'analyse syntaxique au format de Syntex 129

C Liste des prépositions issues de PrepLex 131

D Entrées de LexSchem pour le verbe *accompagner* dans les différents formats disponibles 133

E Liste des SSC de LexSchem 137

F Articles portant sur l'acquisition de classes verbales 139
F.1 Extrait de l'article publié dans la revue TAL en 2010 . . . 139
F.1.1 Méthode . . . 139
F.1.2 Evaluation . . . 141
F.2 Article publié dans les actes de la conférence CoLing en 2010 . . . 144
F.2.1 Introduction . . . 144
F.2.2 French Verb Classes and the Gold Standard . . . 145
F.2.3 Verb Clustering . . . 147
F.2.4 Features . . . 148
F.2.5 Clustering methods . . . 149
F.2.6 Experimental evaluation . . . 150
F.2.7 Evaluation . . . 150
F.2.8 Discussion and Conclusion . . . 154

G Liste des publications liées à la thèse 157

Bibliographie 159

LISTE DES FIGURES

2.1 Stemma pour la phrase « Google a acheté YouTube. » 9

3.1 Extrait de la table 32H du Lexique-Grammaire. 25
3.2 Entrée de LGLex pour le verbe *accompagner* 29
3.3 Entrée 990 de DicoValence 32
3.4 Entrées du Lefff correspondant au verbe *accompagner* 35
3.5 Entrées de TreeLex pour le verbe *accompagner* 38

4.1 Schéma global des méthodes d'acquisition automatique de schémas de sous-catégorisation 46

5.1 Architecture générale d'ASSCi 59
5.2 Analyse syntaxique en constituants 62
5.3 Analyse syntaxique en dépendances 63
5.4 Exemple d'annotation par Syntex 65
5.5 Représentation de l'annotation par Syntex 65
5.6 Variation de la proximité avec TREELEX ($F_{1,5}$) en fonction du seuil 81

6.1 Représentation log-log de la distribution des verbes du corpus LM10 87
6.2 Page d'accueil de l'interface de consultation de LEXSCHEM . . . 94
6.3 Choix du SSC pour le verbe *commander* dans l'interface 94
6.4 Résultats affichés dans l'interface pour l'entrée *commander* / `[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]` 94
6.5 Résultats affichés dans l'interface pour l'entrée `[SUJ:SN, P-OBJ:SP<en+SN>]` / *immigrer* 94
6.6 Extrait du fichier d'évaluation 101

Liste des tableaux

2.1 Liste non-exhaustive des principaux rôles thématiques 14

3.1 Synthèse pour le Lexique-Grammaire 28
3.2 Synthèse pour SynLex et LGLex 30
3.3 Paradigmes valenciels utilisés dans DicoValence (source : manuel d'utilisation de DicoValence) 31
3.4 Cadres de valence pour le verbe *accompagner* 32
3.5 Synthèse pour DicoValence 33
3.6 Synthèse pour le Lefff 34
3.7 Fonctions et catégories utilisées dans TreeLex 37
3.8 Synthèse pour TreeLex 39

4.1 Liste des SSC reconnus par le système Lerner (cette liste est issue de (Brent, 1993)) 49
4.2 Synthèse des méthodes d'acquisition automatique de SSC présentées (les « ? » correspondent à des données non disponibles ; pour *Korhonen (2006)*, il s'agit des chiffres pour le lexique 3) 56
4.3 Synthèse des méthodes d'acquisition automatique de SSC : évaluation (les « ? » correspondent à des données non disponibles ; pour *Korhonen (2006)*, il s'agit des chiffres pour le lexique 3) 56

5.1 Exemple d'annotation par TreeTagger : la colonne de gauche fournit l'étiquette morphosyntaxique tandis que la colonne de droite est le lemme du mot (la liste des étiquettes utilisées par TreeTagger pour le français est donnée en annexe A. 61
5.2 Résultats du repérage des verbes incorrects dans le corpus des 10 ans du Monde (en nombre de lemmes) 67
5.3 Cas traités par ASSCi 69
5.4 Analyse de Syntex correspondant à la proposition « *il les lui reproche au nom du Sartre qu'il aime* » 71
5.5 Résultats obtenus pour quelques combinaisons de seuils lors de l'étalonnage 82

6.1 20 verbes et leurs fréquences dans le corpus LM10 (verbes identifiés à l'aide de l'étiquette donnée par TreeTagger) 88
6.2 Nombre de verbes par nombre de SSC 89
6.3 Entrées de LexSchem pour les verbes travailler et débouler (après filtrage) 89

6.4 Comparaison de LexSchem avec d'autres lexiques de sous-catégorisaiton pour le français (les chiffres entre parenthèses ne sont pas des chiffres communiqués par les auteurs : ils résultent de nos calculs ; DicoVal = DicoValence) 91
6.5 Nombre d'entrées communes et nouvelles dans LexSchem par rapport aux deux ressources de référence (pour les 1583 verbes communs aux 3 ressources) 98
6.6 Données issues de l'analyse manuelle de 150 verbes de LexSchem 101
6.7 SSC correspondants au verbe *confondre* dans LexSchem 103
6.8 Comparaison de LexSchem-EuroParl avec LexSchem-LM10 . 106

F.1 A Levin style gold standard for French 146
F.2 Results for all the features for French (SPEC and K-means) and English (SPEC) 152
F.3 The effect of verb frequency on performance 152

p. 3 examiner : repérer ces constructions est une activité banale pour les utilisateurs de la méthode du lexique-grammaire donc aucune différence avec un autre verbe

Notions de syntaxe et sémantique :

on ne peut pas manipuler des notions complexes sans un minimum de connaissances

Ce ne sont pas comme des données numériques

Pb de définition des ambitions de la thèse : ambiguïté entre l'analyse ~~[illegible]~~ en constituants (délimitation) et l'analyse syntaxico-sémantique tenant compte du sens, de la compositionnalité, des distributions des éléments, de la distinction essentiel/circonstanciel, et des distinctions d'unités lexicales

Certains passages donnent à penser que l'objectif se limite au 1er niveau (ex. fin p. 21) ou inclut le 2e (ex. p. 2). D'autres hésitent entre les deux (ex p. 96)

Pb de définition des objectifs de la thèse :

- description linguistique motivée par la curiosité intellectuelle ?
- applications techniques avec le moins possible de curiosité intellectuelle ?

Ds le 1er cas : objectif non atteint car la description manuelle donne des résultats bcp plus riches

Ds le 2e cas : pour évaluer le degré de succès, il faudrait dire quelles applications sont visées et de quoi elles ont besoin

LISTE DES ABRÉVIATIONS

ACL Association for Computational Linguistics (conférence)

A-OBJ Objet indirect régi par *à*

ATTO Attribut de l'objet

ATTS Attribut du sujet

ANR Agence Nationale de la Recherche

ASSCi Système d'Acquisition de Schémas de Sous-Catégorisation informatisée

Auto Automatique

BNC British National Corpus

CoLing Computional Linguistics (conférence)

DE-OBJ Objet indirect régi par *de*

DGA Direction Générale de l'Armement

EASy Évaluation des Analyseurs Syntaxiques

HPSG Head-driven Phrase Structure Grammar

LADL Laboratoire d'Automatique Documentaire et Linguistique de Paris 7

Lefff Lexique des Formes Fléchies du Français

LFG Lexical-Functional Grammars

LM10 Corpus des 10 ans du journal *Le Monde*

LREC Language, Resources and Evaluation Conference (conférence)

MNCut Multiway Normalized Cut

Nb Nombre

OALD Oxford Advanced Learner's Dictionary

OBJ Objet direct

PCFG Probabilistic Context Free Grammar

PHP PHP : Hypertext Preprocessor

P-OBJ Objet indirect régi par une autre préposition que *à* et *de*

PropSub Proposition subordonnée

RASP Robust Accurate Statistical Parsing

REF Forme pronominale

refl Pronom

SA Syntagme adjectival

SINF Syntagme infinitif

SN Syntagme nominal

SP Syntagme prépositionnel

SSC Schéma de sous-catégorisation

SUJ Sujet

SQL Structured Query Language

TAL Traitement Automatique des Langues

TALC Projet Traitement Automatique des Langues et des Connaissances (`http://talc.loria.fr/`)

TALN Traitement Automatique des Langues Naturelles (conférence)

TLFI Trésor de la Langue Française Informatisé

WOLF Wordnet Libre du Français

XML eXtensible Markup Language

1 INTRODUCTION

1.1 DOMAINE ET MOTIVATIONS

Les informations lexicales, c'est-à-dire les informations se rapportant aux mots et à leurs propriétés ont pris une importance considérable pour le traitement des langues ces dernières années (Danlos, 1988; Laporte, 2000). Il est en effet primordial pour la grammaire d'une langue de savoir comment se combinent les mots, simples ou composés, sur le plan syntaxique comme sur le plan sémantique. La finesse de l'analyse repose finalement moins sur des principes syntaxiques généraux que sur les contraintes propres à chaque élément lexical (Gross, 1975; 1988). Au sein des théories lexicalistes, comme la grammaire lexicale-fonctionnelle (LFG) ou la grammaire syntagmatique guidée par les têtes (HPSG) (Abeillé, 1993), de nombreuses recherches ont été faites pour élaborer un modèle de lexique plus sophistiqué (Bresnan et Zaenen, 1990; Copestake, 1992; Koenig et Davis, 2000). De nouvelles théories visant le développement de lexiques structurés ont alors vu le jour (Levin, 1993; Pustejovsky, 1995).

Parallèlement, la disponibilité de corpus électroniques de grande taille a permis de mener des études sur le comportement linguistique des éléments lexicaux et les propriétés sous-jacentes qui les influencent. En effet, le comportement des mots en contexte est d'autant plus surprenant qu'il peut varier au cours du temps, d'un locuteur à l'autre, voire en fonction de la situation ou du corpus considéré. Pour prendre un exemple simple, le verbe « *essaimer* » est réputé intransitif mais il est largement employé de manière transitive dans le journal *Le Monde* (e.g. « *Cuba a essaimé les effets de son syncrétisme culturel au gré des modes et des engouements* », « *Il n'en a pas moins essaimé son séjour chinois de proclamations optimistes pour l'avenir [du] pays* »). Ce sont au premier chef ces types de comportements non standard (quand on les compare à un lexique général) qui nous intéressent ici [1].

Ce travail de thèse porte donc sur l'acquisition automatique d'informations lexicales à partir de corpus. On s'intéresse plus particulièrement à l'analyse du verbe. Prenons pour exemple une construction typique du verbe *casser* :

$$(Luc)_{\text{ARG1}}\ a\ cassé\ (la\ vitre)_{\text{ARG2}}\ (avec\ un\ ballon)_{\text{ARG3}}.$$

1. Le Trésor de la langue française informatisé enregistre malgré tout quelques emplois transitifs pour *essaimer* dans la langue littéraire. Leur proportion dans *Le Monde* n'en demeure pas moins remarquable.

De manière schématique, on peut distinguer les informations suivantes liées au verbe :

1. Le nombre et la réalisation syntaxique des arguments. Par exemple, « *casser* » peut ici être considéré comme un verbe à trois arguments, le premier (dans la forme canonique du verbe) étant un groupe nominal réalisé à la gauche du verbe (le sujet), le deuxième un groupe nominal à la droite du verbe (l'objet) et le troisième un groupe prépositionnel généralement introduit par *avec*.
2. La grille thématique, qui caractérise le rôle sémantique de chacun des arguments, c'est-à-dire la relation qu'il entretient sur le plan sémantique avec le verbe. Ainsi, dans l'exemple *supra*, l'argument 1 (« *Luc* ») joue le rôle d'agent, l'argument 2 (« *la vitre* ») est le thème et l'argument 3 (« *avec un ballon* ») est l'instrument.

   Notons qu'il n'y a pas de consensus sur la nature et la granularité des rôles thématiques, ni même sur ce qui constitue la grille thématique (Cruse, 1986; Jackendoff, 1990). Ainsi, pour le verbe *casser*, le statut de l'instrumental n'est pas fixé, on l'a considéré ici comme un argument, mais on peut aussi en faire un modifieur suivant le cadre théorique adopté.

Une alternance désigne la relation entre deux réalisations de surface d'un même prédicat, comme *Luc a cassé la vitre avec un ballon* vs *Le ballon a cassé la vitre.* Les alternances ne préservent pas toujours la grille thématique du verbe (cf. *charger le camion de foin* vs *charger le foin dans le camion*, où *camion* passe de thème a but[2]). Les alternances ont été beaucoup étudiées au niveau linguistique (Gross, 1975; Levin, 1993) mais elles restent très difficiles à analyser automatiquement (parmi les rares travaux s'attaquant directement à l'analyse des alternances, voir (McCarthy, 2001)). Les techniques de désambiguïsation sémantique sont encore largement insuffisantes pour distinguer finement les verbes polysémiques (comme *commander*, entre *commander un soda* et *commander à Luc de faire ceci*) (Agirre et Edmonds, 2007).

L'analyse des constructions syntaxiques et l'étiquetage des rôles thématiques ont en revanche suscité de très nombreuses recherches, et nécesssitent des techniques différentes. L'étiquetage des rôles thématiques repose généralement sur une phase d'apprentissage à partir d'un corpus annoté : l'analyseur est ensuite capable d'attribuer des rôles sémantiques en fonction de configurations syntaxiques particulières (sur la question, voir les conférences CoNLL[3] (Stevenson et Carreras, 2009) et (Moreau et al., 2009) pour une expérience sur le français).

L'acquisition de lexiques syntaxiques est un domaine de recherche plus ancien qui s'est développé à partir du début des années 1990 (Brent, 1991; Manning, 1993). Ce courant de recherche repose sur l'idée que les analyseurs syntaxiques non lexicalisés peuvent produire des données relativement structurées de manière massive (en général, à partir de l'analyse de corpus de plusieurs millions de mots), de telle sorte qu'il est ensuite possible de repérer des régularités de comportement et donc d'inférer des connaissances sur les constructions possibles du verbe (c'est-à-dire le nombre et la

2. Sur toutes ces questions, voir le site Sémanticlopédie : `http://www.semantique-gdr.net/dico/`.

3. *Computational Natural Language Learning.*

nature syntaxique des arguments, ce que l'on appelle ici schéma de sous-catégorisation, abrégé en SSC). Ces constructions sont associées "en vrac" aux lemmes verbaux : comme nous l'avons mentionné *supra*, l'état de l'art est encore insuffisant (cf. chapitre 4) pour aller au-delà, faute notamment de techniques de désambiguïsation sémantique suffisamment efficaces. L'information repérée est donc partielle et nécessite un travail manuel de tri et de validation (cf. chapitre 6).

L'hypothèse d'une corrélation entre classes syntaxiques et classes sémantiques a été posée à plusieurs reprises, notamment dans les travaux du LADL dès les années 1970 ((Borillo, 1971) pour les verbes symétriques, (Gross, 1975) pour les verbes psychologiques ou (Boons et al., 1976) pour les verbes locatifs). Nous nous inspirons en outre ici de Beth Levin, qui a proposé de représenter le sens des verbes par le recours à des composants sémantiques (Levin, 1993; Levin et Rappaport Hovav, 2005). C'est en partant de cette notion que Levin arrive à fournir une description systématique des alternances. Elle présume que chaque verbe peut être décrit grâce à un noyau sémantique d'une part, et à un ensemble de composants sémantiques d'autre part. Tandis que le noyau est toujours spécifique à l'unité lexicale, les composants sémantiques sont communs à tous les verbes d'un même groupe sémantique. Les alternances syntaxiques, à leur tour, sont liées à ces composants : les verbes qui appartiennent à un même groupe sémantique (*i.e.* qui partagent les mêmes composants de sens) participent aux mêmes alternances syntaxiques.

Il existe déjà de nombreux dictionnaires à large couverture pour le français (cf. chapitre 3) et il est évident que les méthodes automatiques n'auront ni la richesse ni la précision du résultat d'un travail minutieux mené par une équipe de linguistes[4]. En revanche, il est aisé de montrer que les mots, à l'exemple du verbe « *essaimer* » dans le journal *Le Monde*, ont des comportements variables en fonction du contexte d'usage, et donc du corpus étudié. Il paraît aujourd'hui quasiment inconcevable de repérer ces comportements spécifiques par une analyse manuelle.

Partant de ce constat, les méthodes automatiques ont connu un certain succès car les techniques mises au point permettent 1) de compléter des lexiques existants en mettant le doigt sur des constructions nouvelles ou absentes des dictionnaires, 2) d'obtenir des informations sur la productivités des différentes constructions et 3) d'inférer ensuite des classes de verbes partageant les mêmes alternances, selon l'hypothèse de Levin (Levin, 1993; Schulte im Walde, 2000). Les informations repérées doivent ensuite être validées manuellement puis associées à des entrées lexicales pour produire des ressources vraiment utilisables. Malgré ces limites, la popularité de ce type de méthodes depuis une dizaine d'années montre leur intérêt et leur potentiel pour la mise au point de ressources à large échelle (voir l'expérience décrite dans (Kipper et al., 2008)).

4. L'existence de ressources de qualité (Trésor de la Langue Française, Lexique-Grammaire, *etc.*) à large couverture d'une part, et l'absence de corpus librement disponible d'autre part, a sûrement limité en France l'intérêt pour les méthodes d'acquisition automatique par le passé. Nous pensons toutefois que l'état des techniques est aujourd'hui tel qu'il est intéressant d'y revenir.

1.2 Enjeux

Nous avons identifié trois enjeux importants concernant ce domaine :

1. S'il existe de nombreux travaux pour l'anglais, les techniques d'acquisition automatique de SSC à large couverture pour le français étaient peu étudiées lorsque cette thèse a débuté (la seule étude publiée en 2006 était celle de Paula Chesley et Susanne Salmon-Alt qui ne concerne qu'une centaine de verbes (Chesley et Salmon-Alt, 2006)). Le projet ANR Passage[5] qui a notamment pour objectif l'acquisition de ce type de ressources (Gardent et Lorenzo, 2010) et qui a débuté juste après ce travail de doctorat démontre bien l'émergence et l'importance de cette problématique pour la langue française.
2. Il existe peu d'études approfondies sur l'utilisabilité des systèmes d'acquisition automatique de SSC. L'un des moyens de tester leur utilisabilité est de valider les apports des ressources acquises par rapport aux ressources existantes, en particulier celles qui sont construites par un travail manuel. Il s'agit donc d'utiliser les méthodes d'acquisition tant sur un corpus le plus « général » possible que sur des corpus plus spécialisés et de comparer les informations acquises aux ressources existantes. Les méthodes d'acquisition automatique sont intéressantes si elles permettent d'obtenir des informations nouvelles qui représentent bien des usages de la langue.
3. Il s'agit enfin de déterminer dans quelle mesure il est possible d'inférer des informations sémantiques à partir de données syntaxiques et de vérifier l'hypothèse d'une interface entre syntaxe et sémantique évoquée plus haut. Les SSC acquis à partir de corpus peuvent constituer une base pour l'acquisition de classes « sémantiques ».

Cette thèse va tenter de répondre à ces enjeux en développant l'un des premiers systèmes d'acquisition automatique de SSC à large couverture pour le français. L'acquisition d'un lexique de sous-catégorisation permettra d'évaluer l'utilisabilité de ce système et de vérifier les rapports entre syntaxe et sémantique.

1.3 Plan de la thèse

Dans le premier chapitre (**Description du verbe pour l'élaboration de lexiques**), nous proposons une description du verbe pour l'élaboration de lexiques. Après avoir présenté la notion d'entrée lexicale, nous nous penchons sur les notions de prédicat et de structure argumentale. Nous décrivons ensuite les différentes composantes des schémas prédicatifs : schéma de sous-catégorisation, grille thématique et restrictions de sélection avant de revenir sur la notion d'alternance. Enfin, nous voyons les difficultés qui rendent les contours des schémas prédicatifs flous.

Le deuxième chapitre (**Etat de l'art des ressources existantes pour le français**) fait un état de l'art des principaux dictionnaires électroniques existants pour décrire le verbe en français, qu'ils aient été construits manuellement, semi-automatiquement ou à partir de corpus arborés.

5. http://atoll.inria.fr/passage/

Le chapitre 3 (**Les méthodes d'acquisition automatique de schémas de sous-catégorisation**) présente les méthodes d'acquisition automatique de schémas de sous-catégorisation à partir de corpus bruts. Nous décrivons tout d'abord les méthodes d'évaluation de ces systèmes en distinguant l'évaluation intrinsèque de l'évaluation extrinsèque et en soulignant les problèmes posés par les méthodes d'évaluation généralement utilisées. Nous présentons ensuite un schéma global que semblent partager ces méthodes. Le reste du chapitre fait l'état de l'art des travaux existants : nous présentons d'abord les principaux travaux réalisés depuis le début des années 90 puis les études existantes sur la langue française.

Dans le chapitre 4 (**Acquisition de schémas de sous-catégorisation pour le français**), nous décrivons le système d'acquisition automatique de schémas de sous-catégorisation de verbes à partir de corpus brut que nous avons développé pour le français. Après une présentation des outils d'analyse morphosyntaxique et d'analyse syntaxique de surface utilisés par notre système, nous nous intéressons aux différents modules qui composent ce système d'acquisition.

Nous continuons dans le cinquième chapitre (**Utilisation et évaluation des méthodes d'acquisition automatique de SSC**) par une étude de l'utilisation qui peut être faite des méthodes d'aquisition et des ressources qui en résultent. Une première expérience permet d'acquérir un lexique de sous-catégorisation à large couverture pour le français. L'évaluation de cette ressource montre les limites de l'approche mais aussi ses avantages par rapport aux dictionnaires construits manuellement. Nous mettons en évidence l'intérêt de ces méthodes dans une deuxième expérience qui montre l'adaptabilité des méthodes d'acquisition et par la présentation d'outils pour le paramétrage des ressources.

Le dernier chapitre (**Production de classes de verbes sur la base de leur comportement syntaxique**) présente deux expériences de classification automatique de verbes réalisées à partir du lexique présenté dans le chapitre 5. Ces expériences visent à montrer comment la syntaxe peut servir de marchepied à l'acquisition d'informations sémantiques.

2 Description du verbe pour l'élaboration de lexiques

Les informations lexicales, c'est-à-dire les informations se rapportant aux mots et à leur propriétés, ont pris une importance considérable pour le traitement automatique des langues ces dernières années. Il est en effet primordial pour les systèmes d'extraction d'information (Surdeanu et al., 2003) ou d'analyse syntaxique (Carroll et al., 1998) de connaître la façon dont se combinent les mots, au niveau sémantique et au niveau syntaxique.

Dans le cadre de cette thèse, nous nous intéressons en particulier aux prédicats et à la manière dont ils se combinent avec leurs arguments. Ces informations sont décrites par des schémas prédicatifs.

Nous définissons la notion d'entrée lexicale avant d'introduire les notions de prédicat et de structure argumentale. Nous définissons ensuite le concept de schéma prédicatif et ses différentes notions liées : schéma de sous-catégorisation, grille thématique et restrictions de sélection. Nous présentons ensuite les alternances avant de souligner le caractère flou des contours des schémas prédicatifs.

2.1 La notion d'entrée lexicale

Avant de définir le prédicat et les différents types d'informations lexicales susceptibles de caractériser ses arguments, il faut définir ce que nous entendons par lexique. Un lexique est un ensemble d'entrées lexicales auxquelles on associe des propriétés. Nous définissons donc ici la notion d'entrée et nous tentons d'en délimiter les contours.

Du point de vue de la lexicographie, une entrée lexicale (ou *lexie*) est un ensemble syntaxico-sémantique distinct. C'est donc le sens qui permet de distinguer les entrées lexicales et non leur forme (d'un point de vue morphologique).

Par exemple, les formes *voulait* et *veux* sont caractérisées par une même entrée lexicale (celle qui correspond au verbe *vouloir* et à ses formes fléchies) tandis que les deux formes *volé* dans (2a) et (2b) correspondent à deux entrées distinctes :

(1a) *Il voulait te voir hier.*
(1b) *Je veux aller à l'école.*

(2a) *Elle a volé 5 euros à sa mère.*
(2b) *J'ai volé en deltaplane hier.*

On trouve bien 2 entrées (dont l'une est transitive) pour le mot « *voler* » dans **le Trésor de la langue française informatisé (TLFI)** [1].

Néanmoins, la notion d'entrée lexicale est parfois floue : il est difficile de déterminer à partir de quand une nuance de sens forme une entrée lexicale distincte.

Voici quelques exemples ambigüs qui formeront ou non des entrées lexicales distinctes selon les lexiques :

(3a) *Antoine est tombé de haut quand il a appris la nouvelle.*
(3b) *Jean est tombé de sa chaise.*

(4a) *Elle a cassé le vase de maman.*
(4b) *Il nous casse souvent les pieds.*
(4c) *Il a cassé sa pipe.*

(5a) *La balle roule.*
(5b) *Il roule sa bosse.*
(5c) *Mon frère a roulé pendant des heures pour arriver ici.*
(5d) *Les chanteurs d'autrefois roulaient les "r".*

Les phrases (4c) et (5b) sont des expressions figées dont le sens est assez éloigné du sens du verbe seul : par exemple, *casser sa pipe* signifie mourir et constitue un sens bien différent de *casser*. D'autres phrases ((3a), (4b) et (5c) par exemple), la nuance de sens est moins nette et il est difficile de déterminer si *tomber de haut*, *casser les pieds* et *rouler (en voiture)* constituent des entrées lexicales distinctes de *tomber*, *casser* et *rouler*.

La notion d'entrée lexicale est donc une notion riche sur laquelle nous

1. Le TLFI est un lexique en ligne des mots du français qui sera présenté plus en détail dans le chapitre 3.

reviendrons lorsque nous évoquerons le problème des contours des schémas prédicatifs (section 2.5).

2.2 Prédicat et structure argumentale

Nos travaux concernent les informations lexicales relatives aux prédicats. Le prédicat est l'élément qui met en relation un ou plusieurs autres éléments d'une phrase (ou d'une proposition).

Les manières de représenter les prédicats sont multiples. Nous nous contentons ici d'évoquer les travaux de Tesnière (Tesnière, 1959) où le prédicat est le noyau supérieur qui régit les autres éléments de la phrase. Chez Tesnière, la hiérarchie des connexions entre les éléments de la phrase est représentée sous forme d'un « stemma », c'est-à-dire d'un schéma liant le prédicat à ses arguments sans préciser la relation syntaxique entre eux (voir figure 2.1).

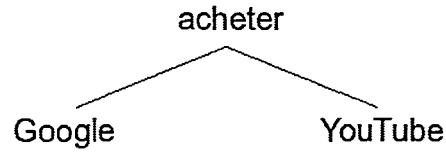


Figure 2.1 – *Stemma pour la phrase « Google a acheté YouTube. »*

Un prédicat peut être un verbe (exemple (6)), un nom (exemple (7)) ou un adjectif (exemple (8)). Dans le cadre de cette thèse, nous nous sommes principalement intéressés aux prédicats verbaux. Néanmoins, pour l'essentiel, le contenu de ce chapitre s'applique à tous les types de prédicats.

(6) *(Julie) (**a acheté**)*$_{\text{PRED}}$ *(un livre) (à Marc).*
(7) *L' (**acquisition**)*$_{\text{PRED}}$ *(de FlickR) (par Yahoo).*
(8) *(Je) suis (**hésitant**)*$_{\text{PRED}}$ *(à lui demander) (de l'aide).*

Les éléments de la phrase qui sont mis en relation par le prédicat forment la structure argumentale de ce prédicat. Nous appellons ces éléments les « compléments » du verbe [2]. Dans la phrase (6), la structure argumentale est formée par les compléments *Julie*, *un livre* et *à Marc*.

Dans les manuels de grammaire (par exemple, dans **le Bon Usage**), le sujet est la plupart du temps exclu de la structure argumentale. En effet, le sujet et le groupe verbal (c'est-à-dire le prédicat verbal et ses arguments) sont considérés comme étant les deux éléments constitutifs de la phrase. Néanmoins, d'un point de vue sémantique, l'élément placé en position sujet est le plus souvent partie prenante de l'action. Il n'y a donc pas lieu de l'exclure de la structure argumentale dans le cadre de notre travail.

2. Nous utilisons dans cette thèse le terme de complément pour tous les éléments reliés au verbe, indépendamment de la distinction argument/modifieur.

Parmi les compléments d'un verbe, on distingue les arguments des modifieurs (Lyons, 1968; Gross, 1975). Les arguments sont obligatoires dans la structure argumentale (ils peuvent être omis en surface mais ils sont toujours présents dans la structure profonde), tandis que les modifieurs sont en général optionnels et plus facilement déplaçables en surface. Nous nous contenterons pour l'instant de cette présentation rapide et nous reviendrons plus loin sur cette distinction entre arguments et modifieurs.

Nous appellons « schéma prédicatif » la description plus formelle qui permet de décrire la structure argumentale d'un prédicat. Selon la théorie, la ressource étudiée (ou éventuellement l'application visée), ces schémas peuvent contenir diverses informations sur le prédicat et ses arguments :

- l'arité du prédicat, c'est-à-dire son nombre d'arguments (par exemple, `3` pour le prédicat *acheter* dans (6)) ;
- le « type » (fonctions et/ou catégories syntaxiques) des arguments (par exemple, `sujet` ou `syntagme nominal` pour *Julie* dans (6)) ;
- le rôle thématique des arguments (par exemple, `agent` pour *Julie* dans (6)) ;
- des restrictions de sélection sur les arguments (par exemple, `+humain` pour *Julie* dans (6)) ;
- les alternances qu'accepte le prédicat (par exemple, l'`effacement de l'argument` *à Marc* dans (6)).

Ces différentes propriétés seront présentées plus en détail tout au long de ce chapitre hormis l'arité qui ne nécessite pas d'explications supplémentaires. Toutefois, la distinction argument/modifieur n'étant pas triviale, la notion d'arité ne peut pas non plus être triviale.

Le schéma prédicatif est donc un objet à géométrie variable. L'examen plus détaillé de chacune de ses composantes potentielles est un préalable pour pouvoir déterminer ce dont nous avons besoin dans nos schémas et ce qui est plus accessoire selon le contexte scientifique et/ou applicatif dans lequel nous nous plaçons.

2.3 Les différentes composantes des schémas prédicatifs

Nous venons de le voir, la structure argumentale d'un prédicat peut être décrite à différents niveaux. Après un bref exemple introductif et un préalable sur la distinction argument/modifieur, nous présentons les concepts fréquemment utilisés pour caractériser la structure argumentale des prédicats : schémas de sous-catégorisation ; grilles thématiques et restrictions de sélection.

2.3.1 Exemple introductif

Considérons les phrases suivantes :

(9a) *Jean donne un livre à Marie.*
(9b) *Jean donne à Marie un livre qu'il a lu récemment.*
(9c) *Jean le lui a donné.*
(9d) *Un livre est donné à Marie par Jean.*

On remarque que l'ordre des éléments de la phrase (ou « réalisation de surface ») est différent pour chacune des phrases bien que ces phrases expriment le même procès.

Par ailleurs, la relation syntaxique entre les arguments et le prédicat reste inchangée hormis dans le cas (9d) (forme passive). Dans la phrase (9d), « un livre » est le `sujet` du verbe alors qu'il est `complément d'objet` dans les phrases (9a), (9b) et (9c).

Cependant, le rôle des éléments du point de vue du sens reste inchangé : dans chacune des phrases, *Jean* donne le livre tandis que *Marie* le reçoit. Dans chacun des cas, le *livre* est l'objet que *Jean* donne à *Marie*, même lorsqu'il se trouve en position `sujet` (phrase (9d)).

Schématiquement, la grille thématique représente la part sémantique du schéma prédicatif alors que son schéma de sous-catégorisation concerne plutôt sa syntaxe. Avant de présenter plus en détail ces notions ainsi que celle de restriction de sélection, il est nécessaire de revenir sur la distinction argument/modifieur.

2.3.2 Distinction argument/modifieur

Classiquement, on différencie les arguments des modifieurs par le caractère obligatoire des arguments au niveau de la structure sémantique profonde (Lyons, 1968; Gross, 1975). Les modifieurs ne font pas partie de la structure argumentale. Ils sont donc facultatifs, plus facilement déplacables, *etc.*. De nombreux travaux ont essayé de dégager les critères linguistiques permettant de distinguer les arguments des modifieurs (Somers, 1984; Pollard et Sag, 1987). (Miller, 1997) fait l'inventaire des principaux critères utilisés : obligatoire *vs.* facultatif ; degré de sélection par le verbe ; thématisation ; *etc.*.

Les exemples suivants montrent la pertinence des critères d'effacement et de déplacement du complément pour différencier arguments (22) et modifieurs (21) :

(21a) *L'abbé Pierre est mort en 2007.*
(21b) *L'abbé Pierre est mort.*
(21c) *En 2007, l'abbé Pierre est mort.*

(22a) *Il a envoyé une lettre à son cousin.*
(22b) *Il a envoyé une lettre.*
(22c) ?*À son cousin, il a envoyé une lettre.*
(22d) ?*Il a envoyé à son cousin une lettre.*

Si on peut sans problème supprimer (21b) ou déplacer (21c) le complément *en 2007* dans la phrase *L'abbé Pierre est mort en 2007.* sans bouleverser la structure sémantique profonde de la phrase, il est plus difficile de déplacer *à son cousin* dans la phrase *Il a envoyé une lettre à son cousin.* ((22c) et (22d)). Néanmoins, dans cet exemple, le critère d'effacement ne permet pas de caractériser le complément *à son cousin* : la phrase (22b) est correcte.

Il faut toutefois souligner que les critères cités plus haut ne font pas l'objet d'un consensus fort. En outre, il n'existe pas de critère linguistique suffisamment discriminant pour différencier à coup sûr, quel que soit le contexte, arguments et modifieurs.

Nous reviendrons sur ce problème dans la section 2.5.2.

2.3.3 Schéma de sous-catégorisation

La « sous-catégorisation » est le phénomène syntaxique qui dénote la tendance des prédicats à imposer à leur entourage des configurations syntaxiques particulières. Ces configurations sont représentées par des schémas (ou cadres) de sous-catégorisation. Dans cette thèse, nous préférerons le terme de sous-catégorisation à celui de valence, parfois utilisé pour décrire le nombre d'arguments d'un prédicat. De plus, nous utilisons le concept de sous-catégorisation uniquement pour caractériser des phénomènes syntaxiques. Or, dans certains travaux, cette notion est utilisée pour décrire aussi bien des phénomènes syntaxiques que sémantiques (par exemple dans (Korhonen, 2002)).

Les schémas de sous-catégorisation (SSC) d'un prédicat décrivent la structure argumentale de ce prédicat en termes syntaxiques. Un SSC est donc composé d'un prédicat et d'une combinaison possible d'arguments pour ce prédicat. Ainsi, l'ensemble des SSC d'un prédicat représente les différents usages syntaxiques possibles de ce prédicat.

Il existe trois représentations syntaxiques possibles des arguments :

- La **fonction** décrit la relation syntaxique entre le verbe et l'argument : sujet, objet direct, objet indirect, attribut du sujet, ...
- La **catégorie syntaxique** du syntagme constituant l'argument exprime sa nature : syntagme nominal, syntagme prépositionnel, syntagme verbal, ...
- La **représentation mixte** mêle les deux représentations précédentes en associant à la fonction de l'argument sa catégorie syntaxique : sujet+syntagme nominal, objet direct+syntagme verbal, ...

L'exemple suivant illustre ces trois représentations (le SSC sous forme réduite est donné entre crochets [3]) :

- Fonctions : *(Julie)*$_{\text{SUJ}}$ ***a donné*** *(un livre)*$_{\text{OBJ}}$ *(à Marc)*$_{\text{P-OBJ}}$.
  `[SUJ, OBJ, P-OBJ]`
- Catégories : *(Julie)*$_{\text{SN}}$ ***a donné*** *(un livre)*$_{\text{SN}}$ *(à Marc)*$_{\text{SP(À)}}$.
  `[SN, SN, SP]`
- Mixte : *(Julie)*$_{\text{SUJ / SN}}$ ***a donné*** *(un livre)*$_{\text{OBJ / SN}}$ *(à Marc)*$_{\text{P-OBJ / SP(À)}}$.
  `[SUJ:SN, OBJ:SN, P-OBJ:SP]`

Dans la phrase « *Julie a donné un livre à Marc.* », *Julie* et *un livre* sont des syntagmes nominaux mais le premier a la fonction sujet tandis que le deuxième est un objet. *à Marc* est un objet prépositionnel de catégorie syntagme prépositionnel.

Cet exemple assez trivial masque le problème de la distinction argument/modifieur dont nous avons parlé plus haut (section 2.3.2) et qui constitue la principale difficulté lorsqu'on travaille sur la sous-catégorisation ver-

3. La signification des abréviations utilisées dans les SSC est donnée dans le tableau 5.3 (chapitre 5).

bale. En effet, certains compléments, présents « en surface » sont des modifieurs et ils ne devraient pas figurer dans le SSC :

(10) *(Barack Obama)*$_{\text{SUJ / SN}}$ ***est devenu*** *(Président des États-Unis)*$_{\text{OBJ / SN}}$ *(le 4 novembre 2008)*$_{\text{MODIFIEUR}}$.
`[SUJ:SN, OBJ:SN]`

(11) *(Le pape)*$_{\text{SUJ / SN}}$ ***rencontrera*** *(les victimes)*$_{\text{OBJ / SN}}$ *(à Malte)*$_{\text{MODIFIEUR}}$.
`[SUJ:SN, OBJ:SN]`

Dans ces 2 phrases, *le 4 novembre 2008* et *à Malte* sont des modifieurs et ils sont par conséquent exclus des SSC.

Néanmoins, il arrive qu'il ne soit pas si simple de déterminer si un complément est un argument ou un modifieur :

(12a) *(Pierre)*$_{\text{SUJ / SN}}$ ***s'est rendu*** *(à Lille)*$_{\text{P-OBJ / SP(À)}}$ *(par l'autoroute A1)*$_{?}$.
`[SUJ:SN, P-OBJ:SP(à), P-OBJ:SP(par)]` ou `[SUJ:SN, P-OBJ:SP(à)]` ?

(13a) *(Les actions)*$_{\text{SUJ / SN}}$ ***ont baissé*** *(de 10%)*$_{?}$ *(la semaine dernière)*$_{\text{MODIFIEUR}}$.
`[SUJ:SN, P-OBJ:SP(de)]` ou `[SUJ:SN]` ?

Il n'existe pas de théorie qui permette de déterminer à coup sûr si *par l'autoroute A1* et *de 10%* sont des arguments ou des modifieurs, certains des critères classiques de distinction se contredisent :

(12b) *Pierre s'est rendu à Lille.*
(12c) *?Pierre s'est rendu par l'autoroute A1 à Lille.*

(13b) *Les actions ont baissé la semaine dernière.*
(13c) **De 10%, les actions ont baissé la semaine dernière.*

par l'autoroute A1 et *de 10%* sont « effaçables » ((12a) et (13a)) mais leur « déplacement » pose problème ((12b) et (13b)).

Les schémas de sous-catégorisation concernent donc les propriétés syntaxiques des arguments d'un prédicat. Les deux autres composantes des schémas prédicatifs (la grille thématique et les restrictions de sélection) caractérisent leurs propriétés sémantiques.

2.3.4 Grille thématique

On appelle grille thématique le schéma constitué par les rôles thématiques des arguments d'un prédicat.

Le rôle thématique d'un argument est une étiquette qui désigne la relation sémantique entre un prédicat et son complément dans un schéma donné (Gruber, 1965; Fillmore, 1968; Jackendoff, 1972; Wilkins, 1988). Il caractérise de façon très générale le rôle de l'argument dans la structure prédicative à un niveau sémantique.Par exemple, dans la phrase *Mon frère a mangé tout le chocolat.*, l'argument *mon frère* est l'`agent` du verbe *manger*.

Cette notion est sujette à discussions : les frontières délimitant les différents rôles sont floues et il n'y a pas de consensus sur le nombre ou la nature

Rôle	Exemple
agent	(Sacha) lance la balle.
patient	Marie attrape (la gomme).
thème	J'achète (un disque) à Marie.
expérienceur	(Léon) admire Paul.
but	Je vous appelle (à propos de l'annonce).
source	Ce vacarme vient (de l'appartement du dessus).
destination	Il va (à Paris) ce week end.
lieu	Le chat est assis (sur le livre).
instrument	Le cuisinier fait sauter ses légumes (avec un wok).
moyen	Nicolas a fait rire Marie (en lui racontant une blague).

TABLE 2.1 – *Liste non-exhaustive des principaux rôles thématiques*

des rôles thématiques (Cruse, 1986; Jackendoff, 1990; Baker, 1997). On peut néanmoins présenter une liste non exhaustive des rôles thématiques les plus fréquemment retenus :

Une grille thématique est simplement une liste de rôles thématiques pour un prédicat donné. Voici quelques exemples de grilles :

Lancer	[agent, patient]	(Zoé)$_{\text{AGENT}}$ a lancé (la balle)$_{\text{PATIENT}}$
Donner	[agent, thème, but]	(Jean)$_{\text{AGENT}}$ donne (le crayon)$_{\text{THÈME}}$ (à Luc)$_{\text{BUT}}$
Acheter	[agent, thème, source]	(Le Brésil)$_{\text{AGENT}}$ a acheté (des Rafales)$_{\text{THÈME}}$ (à la France)$_{\text{SOURCE}}$
Aimer	[expérienceur, patient]	(Il)$_{\text{EXPÉRIENCEUR}}$ aime (sa femme)$_{\text{PATIENT}}$

Lancer prend donc deux arguments dont l'un est `agent` (le « lanceur ») et l'autre `patient` (le « lancé »). Pour le verbe *acheter*, les trois arguments sont `agent`, `thème` et `source`.

Les rôles thématiques sont souvent considérés être à l'interface de la syntaxe et de la sémantique (un agent est de préférence réalisé comme un sujet tandis qu'un patient est plutôt réalisé comme un objet direct). Les exemples précédents montrent bien que s'il existe un lien entre fonction syntaxique et grille thématique, ce lien n'est ni systématique ni évident : remarquons notamment que dans les trois premiers exemples, le sujet du verbe est également `agent` mais dans le cas du verbe *aimer*, il est `expérienceur`.

Enfin, rappelons que la notion de grille thématique n'est pas fixée et les propriétés utilisées pour caractériser les arguments des prédicats dépendent de la théorie ou de la ressource considérée.

VERBNET (Kipper-Schuler, 2005) et FRAMENET (Baker et al., 1998) sont des ressources qui répertorient des informations syntaxiques et sémantiques sur les verbes en anglais : schémas de sous-catégorisation, restrictions de sélection sur les arguments et rôles thématiques. VERBNET permet d'annoter les relations entre le prédicat et ses arguments à l'aide de 23 rôles thématiques tandis que la catégorisation sémantique des arguments dans FRAMENET est plus spécialisée : la ressource fait appel à des « rôles sémantiques » qui sont beaucoup plus précis que les rôles thématiques (par exemple, pour le verbe *acheter*, les arguments seront `acheteur` et `vendeur` (et non `agent` et `source`).

David Dowty propose quant à lui de n'utiliser que deux grands types de rôles qu'il appelle proto-agent et proto-patient (Dowty, 1991) et donne une liste des propriétés qui caractérisent ces deux types. Par exemple, le proto-agent provoque un événement ou le changement d'état d'un autre participant tandis que le proto-patient subit un changement d'état et/ou est affecté par un autre participant.

(14a) *Jean ouvre la porte.*
(14b) *La clef ouvre la porte.*
(14c) *Le vent ouvre la porte.*

Pour les classifications en rôles thématiques « classiques », *Jean* est `agent` dans (14a) et *La clef* est `instrument` dans (14b). Or, les trois éléments en position sujet jouent le même rôle sémantique par rapport au verbe. Pour Dowty, *Jean*, *La clef* et *Le vent* sont tous `proto-agent` tandis que *la vitre* est `proto-patient`.

Les rôles thématiques décrivent les relations sémantiques entre le prédicat et ses arguments ou entre les arguments eux-mêmes mais pas des propriétés intrinsèques des arguments.

2.3.5 Restrictions de sélection sur les arguments

En plus d'informations syntaxiques (sous-catégorisation) et des rôles thématiques, il est possible de décrire des restrictions de sélection sur les arguments d'un prédicat. Les restrictions de sélection sont des propriétés qui permettent de restreindre au niveau sémantique le type de nom accepté par un argument. Par exemple, le premier argument du verbe acheter peut être un humain mais pas un animal (sauf dans certains contextes particuliers) ou un objet inanimé, comme en témoignent les exemples (15a), (15b) et (15c).

(15a) *(Jean) achète un livre à la librairie.*
(15b) *? (Le lapin) a acheté une carotte.*
(15c) * *(La table) a acheté une chaise.*

La phrase (15b) n'est correcte que dans un contexte où *le lapin* a une apparence ou un comportement humain. Remarquons également que la métonymie permet à une organisation (comme par exemple une entreprise) de « prendre » le trait humain :

(15d) *(Google) a acheté YouTube pour 1,65 milliards de dollars.*

Le plus souvent, on représente les restrictions de sélection en utilisant un trait (par exemple, `humain` ou `hum`) qui symbolise une caractéristique sémantique de l'argument. On fait précéder ce trait d'un plus ou d'un moins pour signaler si l'argument est caractérisé par ce trait ou son opposé. Par exemple, `+hum` signifie que l'argument est humain et `-hum` qu'il est non-humain.

Il est nécessaire de distinguer les restrictions de sélection des préférences lexicales. Les préférences lexicales sont l'ensemble des têtes lexicales apparaissant dans une position donnée (par exemple en position sujet ou objet).

En traitement automatique, le terme de restriction de sélection est parfois utilisé pour parler de généralisations sur les têtes lexicales des arguments. Les restrictions de sélection sont alors des classes, c'est-à-dire à des groupes de mots partageant des propriétés sémantiques, auxquelles les arguments du prédicat doivent appartenir. Par exemple, le sujet du verbe *acheter* doit appartenir à la classe des humains ou à celle des organisations.

Pour définir des restrictions de sélection pour un schéma prédicatif donné, on s'appuie sur un inventaire de catégories sémantiques du type animé *vs.* inanimé. Le choix de ces catégories et leur niveau de granularité dépendent du cadre théorique et/ou de l'application visée. Il est également possible de déterminer ces catégories en s'appuyant sur une ressource existante comme par exemple WordNet (Miller, 1995; Fellbaum, 1998). WordNet est une base de données lexicale hiérarchisée organisé en « jeux de synonymes » appelés *synsets*. Ainsi, il est possible d'inférer des restrictions de sélection en s'appuyant sur la hiérarchie de la ressource. Par exemple, dans WordNet, les noms *coffee* (café) et *beer* (bière) ont pour hyperonyme *drinkable* (buvable). Or, le verbe *boire* est le plus souvent rencontré avec des mots de la catégorie *buvable* comme patient. On peut donc déduire de la ressource une restriction de sélection (trait `+buvable`) sur le patient du verbe *boire*.

2.4 Alternances

Une alternance désigne la relation entre deux réalisations de surface d'un même prédicat. Elles mettent en évidence des nuances de comportement tant syntaxiques que sémantiques de verbes et permettent une classification plus fine des prédicats.

Les phrases de l'exemple (9) (voir section 2.3) expriment le même procès mais ont des réalisations de surface différentes :

(9a) (forme canonique) *Jean donne un livre à Marie.*
(9b) (inversion) *Jean donne à Marie un livre qu'il a lu récemment.*
(9c) (pronominalisation) *Jean le lui a donné.*
(9d) (forme passive) *Un livre est donné à Marie par Jean.*

Les alternances décrivent des changements dans la réalisation de la structure argumentale d'un verbe à partir d'une forme canonique (9a). Ces changements peuvent être de différentes natures : déplacement d'argument (comme dans la phrase (9b)), effacement d'argument, transformation de syntagmes prépositionnels en syntagmes nominaux, pronominalisation (comme dans la phrase (9c)), etc.

Les alternances ont été beaucoup étudiées au niveau linguistique (Gross, 1975; Levin, 1993). Dans son ouvrage, Beth Levin définit 79 alternances pour l'anglais (Levin, 1993). Le plus souvent, il s'agit de transformations à partir d'une structure argumentale considérée comme étant la forme canonique (dans notre exemple, la phrase (9a)). Certaines alternances sont très fréquentes tandis que d'autres ne concernent qu'un nombre très limité de verbes.

Par exemple, dans l'alternance `causative/inchoative`[4], le sujet « s'efface » pour laisser sa place à l'objet, comme dans l'exemple suivant :

(16a) *Émilie a cassé la fenêtre.*
(16b) *La fenêtre a cassé.*

Si les alternances impliquent des modifications évidentes dans la syntaxe de la phrase, elles provoquent également des changements sémantiques.

On perçoit cette nuance de sens dans les phrases (16a) et (16b) (dans (16b), on a l'impression que la fenêtre s'est cassée d'elle même) mais l'alternance `locative` met en évidence ce phénomène comme en témoignent les phrases suivantes :

(17a) *Arnaud a chargé le camion d'oranges.*
(17b) *Arnaud a chargé les oranges dans le camion.*

Ici, l'alternance `locative` n'a pas préservé la grille thématique du verbe (dans (17b), *camion* passe de thème à but). De plus, la phrase (17a) implique que le camion soit rempli d'oranges alors que ce n'est pas le cas pour la phrase (17b) : les alternances ne sont donc pas uniquement un phénomène syntaxique.

Les verbes partageant les mêmes alternances semblent partager une certaine proximité sémantique. Levin montre que les alternances, qui décrivent des comportements majoritairement syntaxiques, permettent d'aboutir à une classification « sémantique » des verbes. Ces classes verbales sont variées : verbes de « jet », verbes de contact avec impact, verbes d'interaction sociale, verbes de perception, verbes de désir, verbes de communication, etc.

Pour montrer l'importance des alternances dans la classification des verbes, remarquons tout d'abord que deux verbes partageant les mêmes schémas prédicatifs peuvent ne pas accepter les mêmes alternances. Par exemple, l'alternance `causative/inchoative` est applicable à *casser* mais ne l'est pas à *frôler* :

(18a) *Le ballon casse la vitre.*
(18b) *La vitre casse.*

(19a) *Le ballon frôle la vitre.*
(19b) * *La vitre frôle.*

Aucune information de surface n'est ici disponible pour distinguer le cas (18) du cas (19). C'est donc bien la différence sémantique entre *casser* et *frôler* (en particulier le changement d'état produit par « *casser* » sur la vitre) qui rend possible l'alternance `causative/inchoative`. L'alternance permet de différencier les verbes *casser* et *frôler* et de les classifier plus finement.

Cependant, les alternances restent très difficiles à analyser automatiquement (parmi les rares travaux s'attaquant directement à l'analyse des alternances, voir (McCarthy, 2001)). Les techniques de désambiguisation sémantiques sont encore largement insuffisantes pour distinguer finement les verbes polysémiques (comme *commander*, entre *commander un soda* et *commander à Luc de faire ceci*) (Agirre et Edmonds, 2007).

4. Nous reprenons ici l'appellation de Beth Levin.

2.5 Les contours flous des schémas prédicatifs

Nous avons donné un panorama des représentations possibles de la structure argumentale des prédicats. Il s'agit désormais de mettre l'accent sur le fait que les contours de ces schémas sont difficiles à définir et sujets à débats. Ces frontières floues posent des problèmes considérables lorsqu'on cherche à faire l'inventaire des schémas prédicatifs pour constituer une ressource lexicale. Nous tentons ici de donner un aperçu de la complexité du concept de schéma prédicatif et de ses écueils.

2.5.1 Polysémie et entrées lexicales

Revenons tout d'abord sur la difficulté à constituer des entrées lexicales à partir d'une forme polysémique. En théorie, chaque lexie devrait avoir ses propres schémas prédicatifs. Par exemple, les phrases *L'avion vole depuis deux heures.* et *Il m'a volé ma montre.* correspondent à deux emplois différents de la forme *voler* et correspondent donc chacune à un schéma d'un verbe distinct (et non à deux schémas distincts d'un même verbe).

Par exemple, les phrases suivantes représentent des constructions différentes du verbe *compter* :

(20a) *Luc compte jusqu'à 10.*
(20b) *Pierre compte venir à ton anniversaire.*
(20c) *Julien compte sur toi.*
(20d) *Hélène compte sur ses doigts.*

Les phrases (20a) et (20d) correspondent au sens « concret » du verbe *compter*. Les phrases (20b) et (20c) correspondent à d'autres sens du verbe *compter* : dans (20b), *compter* peut être remplacé par *avoir l'intention de* alors que dans (20c), il fait plutôt référence à la confiance accordée à quelqu'un.

Cependant, nous disposons de peu d'informations en surface pour déterminer le sens du verbe *compter* selon le contexte : par exemple, pour (20c) comme pour (20d), *compter* est suivi d'un complément prépositionnel dont la préposition est *sur*. Ce sont des informations sémantiques plus profondes qui permettent de différencier (20c) de (20d).

Ce problème est encore plus fréquent avec la préposition *à* qui est fréquemment utilisée, tant pour des compléments de lieu (modifieur) que pour des objets prépositionnels, ce qui nous amène à la distinction argument/modifieur.

2.5.2 Retour sur la distinction argument/modifieur

Nous avons présenté les critères classiques de distinction des arguments et des modifieurs dans la section 2.3.2 : effacement ou déplacement du complément, thématisation, etc. Toutefois, ces critères ne font pas l'objet d'un consensus fort. Il n'existe pas de critère linguistique suffisamment discriminant pour déterminer de manière certaine, indépendamment du contexte, si un complément est un argument et modifieur. Certains cas sont difficiles et sujets à débat comme l'exemple (13) rencontré plus haut :

(13a) *Les actions ont baissé de 10% la semaine dernière.*

(13b) *Les actions ont baissé la semaine dernière.*
(13c) **De 10%, les actions ont baissé la semaine dernière.*

Il est difficile de dire si le complément *de 10%* est un argument ou un modifieur. Les critères de déplacement et d'effacement sont contradictoires et le complément a un statut un peu particulier.

Christopher Manning synthétise de façon assez éclairante cette question (Manning, 2003) :

> Il y a des arguments très évidents (sujets et objets) et des modifieurs très évidents (de temps ou de lieu) mais aussi beaucoup de choses entre ces deux pôles. Ces choses sont classées tantôt comme arguments, tantôt comme modifieurs selon les besoins théoriques ou les visées de l'auteur. [5]

Il existerait donc un continuum entre arguments et modifieurs, ce qui permet de mieux représenter le caractère plus ou moins obligatoire des « arguments » des verbes. Il faut désormais trouver un moyen de représenter ce continuum en termes de SSC. La réponse apportée par Manning est de représenter la sous-catégorisation comme une distribution de probabilité sur les arguments des schémas :

> Plutôt que de maintenir une distinction catégorique entre arguments et modifieurs et d'avoir à prendre des décisions binaires à propos de tels cas, nous devrions plutôt essayer de représenter l'information de sous-catégorisation comme une distribution de probabilité sur les arguments, où chaque complément est censé apparaître avec un verbe donné avec une certaine probabilité. [6]

L'information contenue dans ces probabilités est évidemment plus riche que la distinction binaire argument/modifieur. Cette approche probabiliste pose cependant un problème pour l'évaluation de la ressource. En effet, comme nous le verrons au chapitre 6, il devient alors difficile de comparer la ressource avec des ressources où la distinction argument/modifieur est généralement « forte ».

2.5.3 Formes pronominales

L'utilisation du pronom est une spécificité du français. Il s'agit d'un phénomène massif (qui concerne de nombreux verbes) et qui est souvent mal traité par les lexiques. Il est donc important de s'interroger sur la façon de traiter les formes pronominales dans un lexique de schémas prédicatifs. En effet, la même forme de surface *se* + *verbe* peut rendre compte de comportements très variés.

Voici les différents cas possibles d'utilisation de la forme pronominale :

5. There are some very clear arguments (normally, subjects and objects), and some very clear adjuncts (of time and 'outer' location), but also a lot of stuff in the middle. Things in this middle ground are often classified back and forth as arguments or adjuncts depending on the theoretical needs and convenience of the author.

6. Rather than maintaining a categorical argument / adjunct distinction and having to make in/out decisions about such cases, we might instead try to represent subcategorization information as a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability.

- Construction pronominale réfléchie, lorsque le sujet exerce une action sur lui même : *Marie se lève.*

- Construction pronominale réciproque, lorsque les sujets exercent une action chacun sur chacun des autres : *Ses proches se sont recueillis sur sa tombe.*

- Construction pronominale subjective, lorsque le pronom est incorporé au verbe : *Un détenu s'est évadé de la prison de Reims.*

- Construction pronominale passive, lorsque le sujet subit l'action mais que l'agent n'est, en général, pas indiqué : « *Les feuilles mortes se ramassent à la pelle.* »

Il existe des verbes qui admettent plusieurs constructions pronominales. Par exemple, *(s')apercevoir* :

(23a) *Jean s'aperçoit dans le miroir.* (construction pronominale réfléchie)

(23b) *Ils s'aperçoivent sur la place.* (construction pronominale réciproque)

(23c) *On s'est aperçu que le virus gagnait du terrain.* (construction pronominale subjective)

(23c) *La montagne s'aperçoit de loin.* (construction pronominale passive)

Les ambiguïtés sont alors difficiles à lever automatiquement. En effet, il semble impossible de différencier tous ces cas en se basant uniquement sur des indices de surface. Seules des connaissances sémantiques nous permettent de faire la différence entre les différentes formes de constructions pronominales.

Cette multiplicité de cas pour les formes pronominales pose la question de leur place dans un lexique : est-ce que ces formes doivent constituer des entrées lexicales distinctes ou est-ce que la forme pronominale constitue simplement un schéma prédicatif distinct ? La réponse à cette question n'est pas triviale :

- les constructions subjectives devraient constituer une entrée lexicale étant donné qu'il s'agit de la seule forme possible pour ces verbes (*s'évanouir vs. *évanouir*) ou d'une forme dont le sens est totalement distinct de la forme non passive (*s'agir vs agir*) ;
- il est plus difficile de déterminer comment intégrer les constructions réfléchies, réciproques ou passives dans les lexiques : elles sont employées de la même manière (sémantiquement) que l'emploi non pronominal dans certains cas (*Pierre lance la balle à Marie. vs. Les enfants se lancent la balle.*) et d'une manière différente dans d'autres cas (*j'ai trouvé un stylo. vs. le stylo se trouve sur le bureau.*).

Nous verrons au chapitre 3 que ce problème complexe est souvent mal pris en compte dans les ressources actuellement disponibles, même dans les lexiques constitués manuellement par des linguistes.

2.5.4 Verbes supports et expressions figées ou semi-figées

La constitution de ressources lexicales se heurte également à la difficulté à déterminer le statut des verbes supports et des expressions figées ou semi-figées.

Les verbes supports ne remplissent pas la fonction de prédicat de la phrase mais servent de « support » à un nom prédicatif ou à un adjectif prédicatif qui remplit ce rôle (Gross, 1975; 1981; 1998; Mel´cuk, 2004) :

César conduit une attaque contre un fort.

Le verbe *conduire* n'est pas le prédicat de la phrase, *une attaque* remplit ce rôle. Une solution possible est de coder l'« expression » *conduire une attaque* directement dans le lexique.

Les expressions figées sont des suites de mots qui n'admettent que des transformations mineures (articles, temps des verbes, insertions d'adjectifs) (Gross, 1975; 1996) :

Il a pris le taureau par les cornes.

L'expression *prendre le taureau par les cornes* a un sens abstrait et l'insertion d'un nouveau syntagme dans l'expression fait perdre ce sens abstrait à l'expression : *Il a pris le taureau de son père par les cornes.*

Les exemples suivants donnent un aperçu de la variété des cas existants pour les verbes *casser* (expressions (semi-)figées) et *faire* (verbe support et expressions figées)) :

Il a cassé sa pipe.
Ils ont cassé la baraque hier soir.
Il a cassé sa tirelire pour te faire ce cadeau !
Il me casse les pieds avec ses histoires.

Google a fait l'acquisition de YouTube. (verbe support)
Arena a fait l'annonce du licenciement de 93 personnes. (verbe support)
Le parquet a fait appel de la décision du tribunal. (expression figée)
Les prêtres font voeu de chasteté. (expression figée)

Les verbes supports et les expressions figées sont des cas difficiles à classer (Gross, 1988). En effet, il est possible de faire le choix de classer chacune des phrases précédentes dans une entrée distincte : *casser sa pipe*, *casser la baraque*, *casser sa tirelire* et *casser les pieds* peuvent former des entrées lexicales distinctes et avoir chacune leurs schémas prédicatifs. Mais on peut également considérer que ces expressions relèvent de la forme transitive du verbe *casser* et constituent simplement des cas particuliers de l'emploi transitif. Il en est de même pour les 4 phrases qui concernent le verbe *faire*.

Ici encore, une analyse sémantique profonde est nécessaire pour distinguer les différents cas de figure et prendre les meilleures décisions. Cependant, les choix effectués lors de cette analyse varient selon l'observateur et les besoins varient selon l'utilisation qui sera faite de la ressource : certaines applications ont besoin d'un traitement fin des expressions figées et verbes supports alors que dans d'autres cas, ce traitement n'est pas nécessaire.

2.6 Conclusion : vers l'acquisition automatique de schémas prédicatifs

Nous avons présenté différents types d'informations susceptibles de décrire la structure argumentale des verbes en français. Nous avons également souligné les difficultés que l'on rencontre lorsqu'on tente de faire l'inventaire des schémas prédicatifs d'une langue.

Ces difficultés se révèlent encore plus prégnantes dans une perspective d'acquisition automatique de ces informations. C'est pourquoi il est nécessaire de faire des choix quant au type d'information dont nous souhaitons disposer en priorité dans notre ressource lexicale. Nous nous concentrons donc sur l'acquisition de schémas de sous-catégorisation qui semble être un bon point de départ pour l'acquisition de schémas prédicatifs au sens plus large (c'est-à-dire contenant des grilles thématiques et/ou des restrictions de sélection).

Avant de déterminer les contours des schémas de sous-catégorisation que nous souhaitons acquérir, il faut réaliser un état de l'art des ressources existantes et des méthodes d'acquisition automatique de schémas de sous-catégorisation.

Les ressources existantes pour le français

3

Nous avons présenté dans le premier chapitre la description lexicographique du verbe qui sert de base à la constitution de dictionnaires.

On distingue trois types de ressources lexicales : les dictionnaires « papier » (à destination des humains) ; les dictionnaires informatisés (version informatisée des dictionnaires papier, à destination des humains mais qui peuvent parfois être exploités par des machines) et les dictionnaires électroniques (à visée informatique). Dans ces derniers, la description doit être explicite : toutes les informations contenues dans ces dictionnaires sont exploitables par des ordinateurs plutôt que par des humains.

Les lexiques de sous-catégorisation font l'inventaire des différentes constructions possibles pour chaque entrée lexicale. Pour chacune de ces entrées, son (ou ses) schéma(s) prédicatif(s) spécifiant le nombre et le type de ses arguments, et les informations complémentaires qui s'y rapportent (par exemple les lemmes têtes des arguments). De tels lexiques peuvent notamment améliorer les performances d'analyseurs syntaxiques (Carroll et al., 1998) ou de systèmes d'extraction d'information (Surdeanu et al., 2003).

Il existe de nombreuses ressources décrivant les verbes en français et nous nous limitons ici aux dictionnaires informatisés et les lexiques électroniques qui contiennent des informations de sous-catégorisation. Les ressources acquises automatiquement seront présentées conjointement avec leur méthode d'acquisition dans le chapitre 4. Elles sont donc exclues du présent inventaire.

Nous présentons dans ce chapitre un état de l'art des ressources existantes pour le français. Nous faisons également l'état de l'art de l'acquisition de lexiques à partir de corpus arborés dans ce chapitre, dans la mesure où ces travaux requièrent un important travail manuel.

3.1 Le Lexique-Grammaire

Le Lexique-Grammaire est la plus ancienne ressource électronique pour le français qui regroupe des informations sur la syntaxe des verbes, noms et adjectifs dans un format électronique avec description systématique (Gross, 1975; 1994; Leclère, 2005).

Le Lexique-Grammaire est une théorie et une pratique de la description exhaustive des langues, inspirée de la théorie transformationnelle et distributionnelle de Zellig S. Harris (1976). La description du français a d'abord été menée au Laboratoire d'Automatique Documentaire et Linguistique (LADL) par une équipe de linguistes et d'informaticiens dirigée par Maurice Gross depuis la fin des années 1960, et continue d'être maintenue et enrichie par l'équipe informatique-linguistique de l'Institut Gaspard-Monge de l'Université Paris-Est Marne-la-Vallée.

Le Lexique-Grammaire est un dictionnaire syntaxique constitué d'un ensemble de matrices binaires appelées tables. Chaque table regroupe les éléments prédicatifs (verbes, adjectifs, noms) qui partagent la même construction type. Cette construction est une structure de phrase simple, c'est-à-dire « canonique » (sans transformation), qui définit la table. Une phrase simple est définie par le nombre et la nature morpho-syntaxique et sémantique des arguments. Chaque table comprend également un ensemble de propriétés distributionnelles, transformationnelles et sémantiques, que vérifient, ou non, les éléments prédicatifs qui figurent en en-têtes des lignes. Les signes + et − marquent le fait qu'une propriété est vérifiée ou non par un prédicat.

La figure 3.1 est un extrait de la table 32H du Lexique-Grammaire[1]. Chaque colonne de la table représente des propriétés que peuvent prendre les verbes relevant de cette table (notamment les constructions syntaxiques possibles). Chaque ligne correspond à un verbe. Les plus et les moins dans les cases permettent de savoir si tel verbe accepte telle propriété[2].

Le Lexique-Grammaire est généralement présenté comme un lexique syntaxique. Toutefois, lorsqu'on observe les tables, on remarque que certains arguments sont décrits par des traits plus « sémantiques » (par exemple `Nhum` signifie que l'argument est généralement un humain[3]). Les informations codées dans les tables du Lexique-Grammaire ne se limitent donc pas à la syntaxe et concernent également des propriétés proches des restrictions de sélection. Les colonnes qui concernent ces propriétés ont un format du type `N0 =: Nhum`.

Par exemple, les phrases suivantes permettent d'illustrer une partie des informations codées dans la table 32H pour le verbe *accompagner* :

(1a) * *Le poireau accompagne Luc.*
(1b) *Léa est accompagnée par Max.*

1. La table 32H contient le verbe *accompagner* que nous utiliserons pour illustrer les dictionnaires tout au long de ce chapitre.

2. Certaines colonnes sont placées à gauche de l'entrée. Elles contiennent des informations concernant les éléments habituellement situés à gauche du verbe (restrictions de sélection sur le sujet, *se* figé, négation, etc).

3. Nous avons vu dans le premier chapitre que certains éléments non humains (comme par exemple les organisations) pouvaient parfois prendre le trait humain du fait de la métonymie.

N0 =: Nhum	N0 =: Nnr	N0 être V-n	N1 V	Ppv	Ppv =: se figé	Nég	<ENT>	N1 être V-n	N0 V Nhum sur ce point	N1 est Vpp W	N1 =: Nconc	N1 =: Nabs	<OPT>
+	-	-	-	<E>	-	-	abandonner	-	-	-	-	-	La chance abandonne Max
-	+	-	-	<E>	-	-	abasourdir	-	-	+	-	-	Le bruit a abasourdi Max
+	-	-	-	<E>	-	-	abattre	-	-	-	-	-	La police a abattu le truand
+	-	-	-	<E>	-	-	abattre	-	-	+	-	-	Le peuple abattra le tyran
~	~	~	~	<E>	-	-	aborder	~	~	~	~	~	Max a abordé un passant dans la rue
~	~	~	~	<E>	-	-	accabler	~	~	~	~	~	Le témoignage accable l'accusé
+	-	-	-	<E>	-	-	accolader	-	-	-	-	-	Le ministre a accolade le héros
+	-	-	-	<E>	-	-	accompagner	-	-	+	-	-	Max accompagne Léa
+	-	-	-	<E>	-	-	accoster	-	-	-	-	-	Max a accosté une dame dans la rue
+	-	-	-	<E>	-	-	accrocher	-	+	-	-	-	Jean a accroché une nana dans la rue
+	-	-	-	<E>	-	-	accrocher	-	-	-	-	-	Les guérilleros ont accroché les soldats dans le défilé
+	-	-	-	<E>	-	-	acheter	-	-	+	-	+	Max a acheté un député
~	~	~	~	<E>	-	-	achever	~	~	~	~	~	On a achevé les blessés
+	-	+	-	<E>	-	-	administrer	+	-	-	+	+	Ce fonctionnaire administre un grand nombre d'employés
+	-	-	-	<E>	-	-	adopter	-	+	+	+	+	Paul a adopté un petit Hindou
~	~	~	~	<E>	-	-	affamer	~	~	~	~	~	Luc affame (le peuple+la ville)

FIGURE 3.1 – *Extrait de la table 32H du Lexique-Grammaire.*

(1c) * *Max accompagne Léa sur ce point*

Le + à l'intersection des colonnes grisées `N1 est Vpp W` indique que le verbe *accompagner* accepte la forme passive (phrase 1b). De plus, le + dans la première colonne de la table signifie que ce verbe n'accepte qu'un No (sujet de la phrase simple) humain (1a) tandis que le – dans la colonne `N0 V Nhum sur ce point` indique qu'il n'accepte pas cette forme (exemple 1c).

A peu près 5 000 verbes simples graphiquement différents ont été examinés pour le français et donnent lieu à environ 15 000 entrées (an sens lexicographique) réparties dans 60 tables (Leclère, 2005). Chaque entrée est définie par l'ensemble des propriétés syntaxiques vérifiées dans une table particulière. Une entrée correspond donc à une ligne dans une des tables, ce qui signifie qu'une même entrée peut donner lieu à plusieurs constructions[4] (la forme canonique et ses transformations).

Un verbe (au niveau morphologique) a autant d'entrées qu'il a d'usages qui ont été considérés comme distincts. Ces entrées peuvent apparaître dans des tables distinctes ou dans la même table Leclère (2005).

Par exemple, il y a des entrées pour *réaliser* dans les tables 6 et 32A :

(2a) *Luc a réalisé qu'il avait oublié le pain.* (table 6)

(2b) *Georges Stein a réalisé un tableau intitulé « Paris - Le Pont du Carrousel »*[5]. *(table 32A)*

Dans d'autres cas, les deux usages ont la même forme canonique et font donc partie de la même table (2 lignes distinctes pour le même verbe) :

(3a) *Le bureau communique avec le salon.* (table 35S)

(3b) *Jacques communique avec Léa..* (table 35S)

Les unités polylexicales qui fonctionnent comme des prédicats dans une phrase simple sont aussi décrites et sont appelées expressions figées.

Les développeurs du Lexique-Grammaire ont fait le choix de coder de manière assez fine certaines constructions. Par exemple, la table 32R3 regroupe les constructions transitives résiduelles et révèle parfois une classification surprenante qui peut se révéler mal adaptée à un usage automatique (Messiant et al., 2009). Ainsi, pour un verbe très courant comme *travailler*, le Lexique-Grammaire dispose de 15 emplois différents dont quatre appartiennent à la table 32R3. Les quatre emplois du verbe *travailler* classés dans 32R3 sont distingués par leurs objets directs lexicaux :

(4a) *Max travaille la balle.*

(4b) *Max travaille son texte.*

(4c) *Max travaille cette discipline.*

(4d) *Max travaille l'opinion publique.*

4. La notion d'entrée dans le Lexique-Grammaire est donc différente de celle utilisée pour la plupart des autres lexiques électroniques.

5. Exemple tiré de Wikipédia : `http://fr.wikipedia.org/wiki/Pont_du_Carrousel`

Le sens du verbe *travailler* est différent dans chacun de ces cas : on ne peut pas substituer le même ensemble de synonymes au verbe *travailler* et la traduction du verbe vers une langue cible dépend souvent de la sémantique du complément. Cependant, les critères syntaxiques qui différencient ces quatre cas sont minces (il s'agit dans les quatre cas d'une construction transitive simple). Ces exemples sont à la frontière des expressions libres-expressions figées et la nécessité d'une telle « finesse » de distinction en plusieurs emplois d'une construction transitive dépend de l'utilisation qui sera faite du lexique. L'utilisation de classes sémantiques de noms sur les arguments ou de restrictions de sélection plus fines permettrait des généralisations plus proches de l'usage et des nuances de sens observées.

Avec l'augmentation de la taille du lexique, il est relativement difficile de maintenir la cohérence de classification par tables (Messiant et al., 2009). Ainsi, un même emploi risque de se retrouver dans deux classes différentes. Par exemple, la sélection d'une complétive directe par le verbe *savoir* correspond à deux entrées : la construction `N0hum V Que P =` : *Luc sait que Léa est à Paris*, et la construction `N0hum V par N2hum Que P =` : *Luc a su par Max que Léa est à Paris* (classée dans la table 10). Il nous semble qu'il n'y a ici pas lieu de distinguer deux sens du verbe *savoir* et de classer ces deux usages du verbe dans des tables différentes.

Par ailleurs, les classes dites « résiduelles », marquées par R comme 32R3, ont tendance à regrouper tous les exemples qui n'obéissent pas strictement à des critères de classification. Généralement les exemples accumulés dans cette classe attendent d'être reclassés dans d'autres classes. Seule l'amélioration continue du Lexique-Grammaire allégera ce problème.

Enfin, le classement binaire du Lexique-Grammaire (une construction est dans le lexique ou en est absente) ne permet pas de rendre compte de constructions « limites » en français : une construction légèrement douteuse, vieillie ou à l'inverse pas encore stabilisée sera soit exclue totalement du dictionnaire soit mise au même plan que les constructions les plus courantes. C'est notamment le cas d'anglicismes ou d'américanismes comme *(se) baser sur* ou de solécismes comme *pallier à un problème*.

À l'inverse, certains verbes très rares sont présents dans le Lexique-Grammaire mais absents de toutes les autres ressources. Par exemple, on trouve dans le Lexique-Grammaire le verbe *torchecuter* qui donne trois résultats (tous liés au Lexique-Grammaire) lors d'une requête sur Google.

Le Lexique-Grammaire n'est pas limité à la description des verbes simples : les adjectifs et les substantifs prédicatifs sont également classés selon les mêmes principes dans différentes tables.

Une selection des tables du Lexique-Grammaire du français (60%) est mise à disposition[6] sous la licence LGPL-LR. Les tables du Lexique-Grammaire ayant été créées à l'époque où l'utilisation d'un tel type de ressources de manière automatique n'était pas facilement réalisable, bien des informations sont restées implicites, laissant le soin de décodage aux utilisateurs humains. Le format utilisé est par ailleurs difficilement exploitable directement et nécessite un transcodage. Des travaux récents ont tenté de rendre les informations disponibles dans les tables du Lexique-

6. `http://infolingu.univ-mlv.fr`

Ressource	Lexique-Grammaire
Méthode de développement	manuelle
Nombre de verbes	15 000 (5 000 graphiquement différents)
Utilisabilité en TAL	Difficilement exploitable (format peu adapté)
Divers	Répartition des verbes dans des tables qui regroupent les verbes à comportement syntaxique proche

TABLE 3.1 – *Synthèse pour le Lexique-Grammaire*

Grammaire plus exploitables par des applications de traitement automatique des langues (la section 3.2 présente ces travaux).

Le tableau 3.1 fait la synthèse des informations disponibles pour le Lexique-Grammaire.

Le Lexique-Grammaire est l'une des ressources les plus complètes pour la description syntaxique du français. La caractérisation des verbes et de leurs arguments est fine et la construction manuelle des tables garantit une certaine qualité. Néanmoins, l'exploitation des tables par des applications de TAL n'est pas directement possible (même si nous verrons que des travaux sont en cours pour favoriser l'exploitation des informations contenues dans les tables par des systèmes automatiques).

3.2 Les ressources dérivées du Lexique-Grammaire : Synlex et LGLex

Le Lexique-Grammaire est l'une des ressources les plus importantes existantes pour le français. Son principal défaut est que ce dictionnaire n'a pas été conçu pour être utilisé par des applications de TAL. Des études récentes ont tenté de pallier ce problème en proposant des méthodes de transformation du format du Lexique-Grammaire.

3.2.1 Synlex

SynLex (Gardent et al., 2006) est un lexique de sous-catégorisation verbale du français, créé à partir des tables du Lexique-Grammaire, et complété manuellement. Le lexique contient 5 244 lemmes verbaux et 538 schémas de sous-catégorisation qui donnent lieu à 19 127 entrées (paires verbe-schéma[7]). L'objectif est de rendre les informations contenues dans le Lexique-Grammaire plus utilisables pour les applications de traitement automatique des langues.

Néanmoins, SynLex contient uniquement les informations contenues dans les tables du Lexique-Grammaire qui étaient disponibles en 2006 (Synlex n'exploite que la partie des tables disponibles sous licence LGPL-LR).

7. Il ne faut donc pas confondre ces entrées avec les « entrées » du Lexique-Grammaire.

```
ID=V_32H_8

lexical-info=[locs=(),cat="verb",verb=[lemma="accompagner"],
aux-list=(),prepositions=()]

args=(const=[pos="0",dist=(comp=[cat="NP",hum="true",
introd-prep=(),origin=(orig="N0 =: Nhum"),introd-loc=()])],
const=[dist=(comp=[cat="NP",hum="true",introd-prep=(),
origin=(orig="N1 =: Nhum"),introd-loc=()]),pos="1"])

all-constructions=[absolute=(construction="true::N0 V N1"),
relative=(construction="N1 est Vpp W",construction="[passif
par]")]

example=[example="Max accompagne Léa"]
```

FIGURE 3.2 – *Entrée de LGLex pour le verbe accompagner*

3.2.2 LGLex

LGLex est un lexique qui a été acquis à l'aide de LGExtract (Constant et Tolone, 2010). LGExtract est un outil qui génère un lexique à partir des tables du Lexique-Grammaire, en les élaguant de propriétés qui sont difficilement utilisables à l'état actuel de l'analyse syntaxique automatique. Le résultat est disponible sous la licence LGPLLR : `http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Visualisation.html`.

La figure 3.2 représente l'entrée de LGLex correspondant au verbe *accompagner*. La première ligne de l'entrée (`ID=V_32H_8`) donne l'identifiant de la table dont est tirée l'entrée (cette entrée reprend les informations contenues dans la ligne 8 de la table 32H du Lexique-Grammaire, c'est-à-dire l'entrée du Lexique-Grammaire correspondant à *accompagner* dans la table 32H (voir section 3.1)). La deuxième ligne présente des informations lexicales sur le verbe alors que la troisième ligne donne les informations détaillées sur les arguments. Les constructions possibles du verbe sont données par la ligne `all-constructions`. Ici, le verbe *accompagner* peut prendre les constructions `N0 V N1` (forme canonique : *Max accompagne Léa.*) et `N1 est Vpp W` (construction passive introduite par la préposition *par* : *Léa est accompagnée par Max.*). Le dernier champ de l'entrée est un exemple qui illustre l'entrée (ici *Max accompagne Léa.*).

LGLex est également disponible au format Lefff : LGLex-Lefff[8] [9] (Sagot et Tolone, 2009).

SynLex et LGLex ont rendu les informations disponibles dans le Lexique-Grammaire plus aisément exploitables par les applications de traitement automatique des langues. Toutefois, ces deux ressources n'ont pas encore atteint la couverture du Lexique-Grammaire et une certaine perte d'informations a été concédée lors de leur acquisition. Le tableau 3.2 fait la synthèse de ces deux ressources lexicales.

8. `http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Visualisation.html`

9. le Lefff est présenté à la section 3.4

Ressource	SynLex	LGLex
Méthode de développement	à partir du LG	à partir du LG
Nombre de verbes	5 244	5694
Utilisabilité en TAL	Exploitable	Exploitable

TABLE 3.2 – *Synthèse pour SynLex et LGLex*

3.3 DicoValence

DicoValence (van den Eynde et Mertens, 2006) est un dictionnaire syntaxique construit manuellement dans le cadre méthodologique de l'Approche Pronominale (van den Eynde et Blanche-Benveniste, 1978). Il s'agit du successeur du lexique PROTON, dictionnaire de valence des verbes français (comprenant 8 600 entrées pour 3 700 infinitifs) réalisé pendant la période 1986-1992 (van den Eynde et Mertens, 2003). Le lexique et un manuel détaillé sont disponibles en ligne [10].

DicoValence répertorie les schémas de sous-catégorisation (appelés cadres valenciels dans DicoValence) de 3 738 verbes simples du français, répartis en 8 313 entrées (c'est-à-dire couples verbe-schéma).

Les schémas comprennent le nombre et la nature des compléments valenciels du verbe avec leur fonction syntaxique. De plus, le dictionnaire fournit certaines restrictions sélectionnelles, certaines formes de réalisation (pronominales, phrastiques) des termes, la possibilité d'employer le SSC à la forme passive, et éventuellement des liens avec d'autres cadres du verbe.

Contrairement aux lexiques qui se veulent indépendants des théories linguistiques, DicoValence s'appuie sur l'Approche Pronominale (van den Eynde et Blanche-Benveniste, 1978). Cette approche repose sur l'hypothèse que les pronoms révèlent des différences au sein de constructions de surface en apparence comparables. Ainsi, *je donne un livre à Jean* se pronominalise en *je lui donne un livre* (et non en **je donne un livre à lui*) tandis que *je pense à Marie* se pronominalise en *je pense à elle* (et non en **je lui pense*), d'où la différence de statut entre les compléments en « *à* » révélée par leurs propriétés de pronominalisation. De plus, les pronoms constituent des « éléments de référence minimale » : ils forment un ensemble fermé, ce qui permet de limiter le nombre de combinaisons possibles sur lesquelles porter un jugement pour établir la valence d'un prédicat.

Ainsi, la représentation des informations de sous-catégorisation dans DicoValence a les propriétés suivantes :

- Pour chaque place de valence (ou « paradigme »), le dictionnaire précise le paradigme de pronoms qui y est associé (c'est-à-dire l'ensemble des pronoms possibles dans une position argumentale donnée) et qui couvre leurs lexicalisations possibles « en intension » (c'est-à-dire les formes non pronominales possibles pour cet argument) ;
- la délimitation d'un cadre de valence (appelée « formulation ») repose sur la configuration des paradigmes pronominaux, mais également sur les autres propriétés de construction associées à cette configuration (forme passive, ...).

10. http ://bach.arts.kuleuven.be/dicovalence/

Po	correspond *grosso modo* au sujet de la grammaire scolaire
P1	correspond *grosso modo* à l'objet direct de la grammaire scolaire
P2	correspond *grosso modo* à l'objet indirect de la grammaire scolaire ; les formes non clitiques présentent la préposition *à*
P3	correspond *grosso modo* à l'objet indirect de la grammaire scolaire ; les formes non clitiques présentent la préposition *de* ou *à*
PL	paradigme locatif
PL_PR	préposition(s) fixe(s) ou prototypiques du paradigme PL
PPL_DEV	réalisation du locatif : paradigme de pronoms pour une préposition donnée
PDL	paradigme délocatif
PDL_PR	préposition(s) fixe(s) du paradigme PDL (*de, à partir de, depuis*)
PM	paradigme de manière
PM_PR	préposition(s) fixe(s) du paradigme PM
PMi	paradigme de manière « interne » (*mes efforts se sont avérés inutiles*)
PT	paradigme de temps
PT_PR	préposition(s) fixe(s) du paradigme PT
PQ	paradigme de quantité
PQ_PR	préposition(s) fixe(s) du paradigme PQ
PP	paradigme prépositionnel
PP_PR	préposition(s) fixe(s) du paradigme PP)
AE	adjoint évaluatif (*un vêtement qui tombe bien*)
PX	paradigme sans proportionalité (*on l'a élu président*)

TABLE 3.3 – *Paradigmes valenciels utilisés dans DicoValence (source : manuel d'utilisation de DicoValence)*

Le tableau 3.3 présente la liste des différents paradigmes valenciels utilisés dans DicoValence [11].

Un exemple d'entrée de DicoValence pour le verbe *accompagner* est donné dans la figure 3.3. Cette entrée (l'entrée 990 de DicoValence) correspond au cadre `P0 P1 (PL)` pour le verbe *accompagner* (dans le sens de *to accompany*). Ce SSC correspond à un emploi transitif éventuellement complété d'un complément de lieu (`PL`), comme par exemple dans *Luc a accompagné Léa à Lyon.* Les champs `P0`, `P1` et `PL` donnent les différentes pronominalisations possibles de chacun des arguments.

Les cadres de valence attribués au verbe *accompagner* sont donnés dans le tableau 3.4.

Des traductions en anglais et néerlandais (champs `TR_EN` et `TR_DU`) permettent de distinguer des traductions différentes d'un même lemme verbal, donc des sens différents : la traduction permet de repérer des entrées diffé-

11. Pour des informations sur les notions syntaxiques manipulées, nous renvoyons à l'ouvrage de référence « Pronom et syntaxe » (Blanche-Benveniste et al., 1984) ou à la synthèse dans (van den Eynde et Mertens, 2003), et aux études citées dans ces publications.

```
VAL$ accompagner: P0 P1 (PL)
VTYPE$ predicator simple
VERB$ ACCOMPAGNER/accompagner
NUM$ 990
EG$ ils l'ont accompagné à sa maison
TR_DU$ begeleiden, vergezellen, meegaan (met),
    (weg)brengen
TR_EN$ accompany (to)
P0$ qui, je, nous, elle, il, ils, on, celui-ci,
    ceux-ci
P1$ qui, te, vous, la, le, les, se réc., en Q,
    celui-ci, ceux-ci, l'un l'autre
PL$ 0, où, y, là, ici, là-bas
RP$ passif être, se faire passif
```

FIGURE 3.3 – *Entrée 990 de DicoValence*

Id.	Verbe	Cadre	Exemple
990	accompagner	Po P1 (PL)	*ils l'ont accompagné à sa maison*
1000	accompagner	Po P1 P3	*elle m'a demandé d'accompagner le plat de beaucoup de légumes*
1010	accompagner	Po P1	*une sauce piquante accompagne tous les plats*
1020	accompagner	Po P1 P2	*un petit orchestre accompagnait le chanteur*
1030	accompagner	Po P1	*les événements forts de la jeunesse vous accompagnent jusqu'à la mort*
1040	s'accompagner	Po P3	*une éruption volcanique s'accompagne-t-elle d'un tremblement de terre ?*

TABLE 3.4 – *Cadres de valence pour le verbe accompagner*

Ressource	DicoValence
Méthode de développement	Manuelle
Nombre de verbes	3 738
Utilisabilité en TAL	Nécessite une transformation de format
Divers	Basé sur l'Approche Pronominale

TABLE 3.5 – *Synthèse pour DicoValence*

rentes, dans la mesure où la notion d'entrée est aussi fondée sur la sémantique du verbe.

Par exemple, les entrées 1010 et 1030 semblent correspondre au même cadre mais concernent des sens différents du verbe *accompagner*. L'argument `P1` se pronominalise en *vous* pour l'entrée 1030 (5a) mais pas pour l'entrée 1010 (5b) et il se pronominalise en *ceci* pour l'entrée 1010 (6a) mais pas pour l'entrée 1030 (6b) :

(5a) *les événements forts de la jeunesse vous accompagnent jusqu'à la mort*
(5b) * *une sauce piquante vous accompagne*

(6a) *une sauce piquante accompagne ceci*
(6b) * *les événements forts de la jeunesse accompagnent ceci jusqu'à la mort*

Ces différences de pronominalisation permettent de distinguer ces deux cas, ce qui souligne l'intérêt de l'approche pronominale.

Le tableau 3.5 résume les informations concernant DicoValence. DicoValence est l'une des ressources syntaxiques les plus complètes et précises concernant les verbes les plus courants du français. Si le socle théorique que constitue l'Approche Pronominale permet au dictionnaire d'être cohérent et robuste, le format particulier de la ressource qui résulte de cette théorie rend son utilisation difficile. En effet, certains éléments cruciaux ne sont pas formalisés explicitement ou pas systématiquement (de nombreuses infinitives introduites par *à* ou *de* sont justes mentionnées en tant que paradigme `P2` ou `P3` (complément introduit par *à* ou *de*) et il faut lire la ligne `P2` ou `P3` de l'entrée pour savoir que le complément peut être une infinitive), et des transformations de format délicates sont nécessaires. Toutefois, de telles transformations peuvent faire baisser la précision de la ressource et induire des pertes d'informations. De plus, DicoValence ne contient qu'un nombre très limité d'informations sémantiques sur le verbe et ses arguments.

3.4 LE LEFFF

Le Lefff (Lexique des Formes Fléchies du Français) est un lexique morphologique et syntaxique à large couverture, disponible sous licence libre LGPL-LR [12] (Sagot et al., 2006; Sagot, 2010).

Les informations morphologiques et syntaxiques contenues dans le Lefff résultent de divers travaux :

12. `http://alpage.inria.fr/~sagot/lefff.html`

Ressource	Lefff
Méthode de développement	Mixte
Nombre de verbes	6 825
Utilisabilité en TAL	Exploitable
Divers	Fusion de ressources (Lexique-Grammaire, DicoValence, ...)

TABLE 3.6 – *Synthèse pour le Lefff*

- acquisition automatique (avec validation manuelle) à l'aide de techniques statistiques appliquées sur des corpus bruts (Clément et al., 2004) ;
- acquisition automatique (avec validation manuelle) d'informations syntaxiques (Sagot, 2006) ;
- correction et ajout manuel ou guidé par des techniques automatiques (Sagot et de La Clergerie, 2006) ;
- ajout d'informations contenues dans d'autres ressources, notamment le Lexique-Grammaire et DicoValence (Sagot et Danlos, 2006; Danlos et Sagot, 2008; Sagot et Tolone, 2009).

Des classes de verbes partageant les mêmes schémas ont été construites semi-automatiquement et ont été validées manuellement. Les schémas ont ensuite été projetés sur l'ensemble des verbes partageant la même classe. Le lexique comprend dans sa version actuelle 6 825 lemmes verbaux.

La figure 3.4 est un extrait du Lefff relatif aux informations de sous-catégorisation pour le verbe *accompagner*. Le SSC de la troisième entrée présentée ci-dessus est : `<Suj:cln|sn, Obj:(cla|seréc|sn), Loc:(y|loc-sn)>`. Il s'agit d'un schéma à trois compléments dont le premier est le sujet (qui peut être clitique nominal ou syntagme nominal), le deuxième est l'objet (qui peut être un clitique accusatif, un *se* récursif ou un syntagme nominal) et un complément locatif (le pronom *y* ou syntagme nominal locatif). Cette entrée correspond à l'entrée `990` de DicoValence dont l'exemple est *ils l'ont accompagné à sa maison.*

Le tableau 3.6 synthétise les informations disponibles à propos du Lefff. Le Lefff est de plus en plus un lexique constitué par la fusion de dictionnaires existants. Le format du lexique le rend exploitable par des applications de TAL (une étude a montré qu'il pouvait être utilisé par un analyseur syntaxique (Sagot et Tolone, 2009)).

3.5 D'AUTRES DICTIONNAIRES ÉLECTRONIQUES POUR LE FRANÇAIS

Nous présentons désormais plus succintement des dictionnaires syntaxiques qui ne sont pas disponibles ou qui sont plus difficilement utilisables que les lexiques présentés jusqu'ici.

3.5.1 Le dictionnaire syntaxique des Verbes Français

Le dictionnaire syntaxique des Verbes Français créé par Dubois et Dubois-Charlier (1997) est une classification sémantico-syntaxique des

```
accompagner     v-er:std
100;Lemma;v;<Suj:cln|sinf|sn,Obj:(cla|sn)>;cat=v;
%actif,%passif,%ppp_employé_comme_adj
# D__1030 L__1 M__1 --- DVex:
les événements forts de la jeunesse vous accompagnent jusqu'à
la mort

accompagner     v-er:std
100;Lemma;v;<Suj:cln|sn,Obj:(cla|sn)>;cat=v;
%actif,%passif,%ppp_employé_comme_adj,%se_moyen
# D__1010 L__1 M__1 --- DVex:
une sauce piquante accompagne tous les plats

accompagner     v-er:std
100;Lemma;v;<Suj:cln|sn,Obj:(cla|seréc|sn),Loc:(y|loc-sn)>;cat=v;
%actif,%passif,%ppp_employé_comme_adj
# D__990 L__1 M__1 --- DVex:
ils l'ont accompagné à sa maison

accompagner     v-er:std
100;Lemma;v;<Suj:cln|sn,Obj:cla|sn,Objde:en|de-sn>;cat=v;
%actif,%passif,%ppp_employé_comme_adj,%se_moyen
# D__1000 M__1 --- DVex:
elle m'a demandé d'accompagner le plat de beaucoup de légumes

accompagner     v-er:std
100;s'Lemma;v;<Suj:cln|sn,Objde:en|de-sn>;@pron,cat=v;
%actif
# D__1040  --- DVex:
une éruption volcanique s'accompagne-t-elle d'un tremblement
de terre?

accompagner     v-er:std
100;Lemma;v;<Suj:cln|sn,Obj:cla|seréc|sn|seréfl,Objà:y|à-sn>;cat=v;
%actif,%passif,%ppp_employé_comme_adj
# D__1020 M__1 --- DVex:
un petit orchestre accompagnait le chanteur
```

FIGURE 3.4 – *Entrées du Lefff correspondant au verbe accompagner*

verbes manuellement construite par ces deux linguistes, dont les principes sont proches de ceux du Lexique-Grammaire. On compte dans ce dictionnaire 12 130 verbes et 25 610 entrées (chaque entrée correspond à un couple verbe - schéma de sous-catégorisation). La ressource est disponible sur le site internet du laboratoire MoDyCo [13].

Toutefois, le dictionnaire est difficilement exploitable en l'état par des applications de TAL, notamment en raison de son format.

3.5.2 LexValf

Le projet LexValf, développé à l'Université de Provence, a comme objectif la construction d'un dictionnaire de valence verbale pour le français [14] (Gala et Valli, 2005; Salkoff et Valli, 2005). A chaque entrée sont associés un ensemble de restrictions lexicales et grammaticales et des indices sur la fréquence d'usage tirée d'informations issues du Web. Les principes de base sont ceux de grammaire en chaîne (Salkoff, 1979). Un accent important est mis sur la formalisation des données (le lexique ayant vocation à être utilisé par des applications de TAL). Le dictionnaire comprend les entrées de 975 verbes parmi les plus fréquemment utilisés du français. L'objectif du projet à terme est de réunir les entrées de plus de 6 000 verbes. Étant donné l'ampleur de la tâche, les constructions à verbe support, les constructions verbales figées et les collocations sont exclues du lexique. Diverses ressources sont utilisées pour constituer LexValf : le Lexique-Grammaire (Gross, 1975), des dictionnaires usuels (Grand Robert, Petit Robert ou le Trésor de la Langue Française) ou encore Les Verbes Français (Dubois et Dubois-Charlier, 1997). Ce lexique n'est pas disponible pour l'instant [15].

3.5.3 Trésor de la Langue Française informatisé

Le Trésor de la Langue Française informatisé (TLFI) est la version informatisée du Trésor de la Langue Française (Dendien et Pierrel, 2003). Il s'agit du plus grand dictionnaire de langue française en 16 volumes réalisé par l'Institut National de la Langue Française (INaLF) entre le début des années 1960 et le milieu des années 1990. Initialement conçu pour être édité uniquement sous forme papier, ce dictionnaire est désormais diffusé sur Internet. Même si cette ressource est parfois employée comme un dictionnaire électronique, il s'agit en fait de la version informatisée d'un dictionnaire papier. En conséquence, la description linguistique n'est ni complètement systématique, ni complètement formalisée. En outre, le TLFI comporte des informations syntaxiques limitées qui sont difficilement exploitables dans leur structuration actuelle.

13. http://www.modyco.fr/index.php?option=com_content&view=article&id=1756&Itemid=19

14. http://www.lif.univ-mrs.fr/IMG/html/LEXVALF.html

15. Une interface de consultation du lexique en ligne est protégée par mot de passe.

Fonction	Catégories
SUJ	NP, VPinf, Ssub, VN
OBJ	NP, AP, VPinf, VN, Sint, Ssub
DE-OBJ	VPinf, PP, Ssub, VN
A-OBJ	VPinf, PP, VN
P-OBJ	PP, AdP, VN, NP
ATO	Srel, PP, AP, NP, VPpart, VPinf, Ssub
ATS	NP, PP, AP, AdP, VPinf, Ssub, VPpart, Sint, VN

TABLE 3.7 – *Fonctions et catégories utilisées dans TreeLex*

3.6 TreeLex, un lexique de sous-catégorisation acquis à partir d'un corpus arboré

Depuis le début des années 2000, une autre approche de constitution de ressources lexicales a été explorée. Des travaux d'acquisition automatique à partir de corpus arborés ont été réalisés pour diverses langues (Sarkar et Zeman, 2000; Marinov, 2004; O'Donovan et al., 2005). Ceux-ci reposent sur des corpus arborés, c'est-à-dire des corpus annotés syntaxiquement à la main ou dont l'analyse a été validée par des linguistes. Le système parcourt les arbres syntaxiques et extrait les arguments des verbes pour associer ces constructions à des SSC. Étant donné l'importance du travail manuel requis dans ce type de travaux, nous préférons présenter ces travaux dans ce chapitre plutôt que dans le chapitre suivant, consacré aux techniques d'acquisition à partir de corpus bruts.

Une étude de ce type a été réalisée récemment sur le français par Anna Kupść (Kupść et Abeillé, 2008b). TreeLex (Kupść, 2007; Kupść et Abeillé, 2008a) est un lexique de sous-catégorisation verbale pour le français contemporain extrait automatiquement du corpus arboré de Paris 7 (Abeillé et al., 2003). Le corpus arboré du français est composé d'extraits d'articles du journal *Le Monde* et contient un million de mots. Il a été annoté par 14 catégories lexicales et 12 types de syntagmes. Le système d'acquisition n'exploite que la partie du corpus qui contient les informations fonctionnelles (15000 phrases, soit environ 300 000 mots). La première étape de l'acquisition consiste à parcourir les arbres et à en extraire les schémas. Les formes actives et passives sont normalisées (les formes passives sont passées à la forme active) et les éléments des schémas sont ordonnés.

Cette méthode a permis d'extraire TreeLex, un lexique de sous-catégorisation pour environ 2 000 lemmes verbaux et 180 SSC (moyenne de 2,09 schémas par lemme).

Les étiquettes utilisées pour les éléments des SSC sont approximativement les mêmes que celles du corpus arboré de Paris 7. Le tableau 3.7 fait l'inventaire des différentes fonctions et catégories reconnues dans TreeLex.

La figure 3.5 regroupe les informations contenues dans TreeLex pour le verbe *accompagner*. On retrouve les principales constructions du verbe *accompagner* illustrées par les phrases suivantes :

(7a) `SUJ:NP, OBJ:NP` : *Léa accompagne Luc.*

```
===accompagner (frames: 4; all verbs: 28)
   SUJ:NP, OBJ:NP (13)
   SUJ:NP, OBJ:NP, DE-OBJ:PP (7)
   SUJ:NP, DE-OBJ:PP, refl:CL (7)
   SUJ:NP, P-OBJ:PP, refl:CL (1)
```

FIGURE 3.5 – *Entrées de TreeLex pour le verbe accompagner*

(7b) `SUJ:NP, OBJ:NP, DE-OBJ:PP` : *Pierre accompagne son fromage de vin.*

(7c) `SUJ:NP, DE-OBJ:PP, refl:CL` : *La hausse de l'activité s'accompagne d'une hausse des valeurs.*

(7d) `SUJ:NP, P-OBJ:PP, refl:CL` : *Une fois cuit le pinon est servi chaud et s'accompagne avec le reste de la sauce tomate.*

Récemment, les schémas de sous-catégorisation des adjectifs prédicatifs ont été ajoutés à TreeLex (27 schémas différents pour 271 adjectifs) (Kupść, 2008).

TreeLex est une ressource disponible pour le français [16] qui a l'avantage d'avoir un format assez standard. Néanmoins, la ressource souffre de la taille du corpus d'acquisition ce qui a pour conséquence la non exhaustivité du lexique. De plus, l'expérience est assez peu reproductible pour des langues de spécialité étant donné les contraintes liées à l'établissement de corpus arborés de taille suffisante pour divers domaines.

Le principal avantage des méthodes d'acquisition à partir de corpus arborés est qu'elles s'appuient sur des données peu bruitées, ce qui permet d'obtenir des résultats plutôt satisfaisants en termes de précision. En outre, la distinction argument/modifieur est spécifiée dans les annotations du corpus (par exemple, dans TreeLex, il existe une fonction `MOD` pour les modifieurs). Le problème de la distinction argument/modifieur n'a donc pas la même ampleur pour les systèmes d'acquisition à partir de corpus arboré que pour les systèmes d'acquisition automatique à partir de corpus brut.

Ce travail préalable garantit donc la qualité des annotations contenues dans le lexique et donc des informations acquises à partir de ces annotations. Cependant, même si l'on peut aussi parler d'acquisition dans ce cas, il s'agit en fait d'une approche très différente des expériences d'acquisition automatique à partir de corpus brut. Les corpus arborés sont excessivement rares : la méthode est donc peu portable et ne permet pas de traiter du corpus « tout venant ». En effet, les corpus arborés sont des corpus annotés à la main ou au moins dont l'analyse syntaxique a été validée manuellement. Toute nouvelle étude demande d'annoter un nouveau corpus à la main, ce qui est long et coûteux. Ces méthodes se distinguent donc des travaux d'acquisition automatique à partir de corpus brut parce qu'elles requièrent un important travail manuel pour constituer le corpus, ce qui les rend peu reproductibles. De plus, le format du corpus arboré est rigide et dépend le plus souvent de la théorie qui sous tend son annotation. Enfin, la taille limitée des corpus arborés peut poser des problèmes quant à la qualité du

16. Les différentes version de TreeLex sont disponibles sur le web : `http://erssab.u-bordeaux3.fr/spip.php?article150`.

Ressource	TreeLex
Méthode de développement	Acquis à partir d'un corpus arboré
Nombre de verbes	2 000
Utilisabilité en TAL	Exploitable
Divers	Corpus arboré de Paris 7

TABLE 3.8 – *Synthèse pour TreeLex*

rappel obtenue par les méthodes décrites ci-dessus : il y a naturellement une plus grande probabilité qu'une construction soit manquante dans un corpus d'un million de mots que dans un corpus de 200 millions de mots). La comparaison de notre approche avec **TreeLex** dans le chapitre 6.2 vérifiera cette intuition.

Il convient donc de distinguer l'acquisition à partir de corpus arboré de l'acquisition à partir de corpus brut et de ne pas comparer ces deux approches entre elles : l'acquisition à partir de corpus arborés permet d'obtenir des ressources d'une certaine précision le plus souvent au détriment du rappel. Le principal avantage des systèmes d'acquisition à partir de corpus brut sur ces méthodes est qu'ils permettent de faire émerger de nouveaux comportements de la langue, en fonction du type de corpus ou du domaine étudié (voir section 6.3.1).

Le tableau 3.8 résume les informations disponibles à propos de TreeLex.

3.7 CONCLUSION

Nous avons vu qu'il existe aujourd'hui de nombreux lexiques contenant des informations de sous-catégorisation pour le français. Ces ressources sont très hétérogènes, tant au niveau de leur format, de leur couverture ou des informations qu'elles contiennent (nombre de verbes et de schémas, caractéristiques des schémas, etc).

Ces lexiques ont été construits manuellement ou semi-automatiquement. Depuis les années 90, d'autres ressources ont été acquises par des méthodes d'acquisition automatique. Le chapitre suivant présente les méthodes d'acquisition automatique de ressources de sous-catégorisation à partir de corpus.

4

Les méthodes d'acquisition automatique de schémas de sous-catégorisation

La construction de dictionnaires de sous-catégorisation de verbes a été l'objet de nombreux travaux. Ce phénomène est particulièrement vrai pour le français, comme en témoigne la variété des lexiques présentés dans le chapitre précédent. Certains de ces lexiques ont été créés avant l'essor de l'informatique et leur format n'est pas adapté aux applications de TAL. Or, le développement du domaine du traitement automatique des langues a rendu de plus en plus nécessaire la disponibilité de lexiques utilisables par les machines.

Les travaux en TAL concernent souvent des sous-domaines ou des corpus de spécialité. Ils nécessitent donc des ressources adaptées à ces domaines. L'adaptation « à la main» des ressources existantes à un domaine particulier requiert un travail long et fastidieux alors qu'un système d'acquisition automatique à partir de corpus brut pourra facilement s'adapter (moyennant la disponibilité d'un corpus pour le domaine et des modifications mineures du système, voir la section 6.3.1 pour plus de détails sur ce sujet). De plus, les applications de traitement automatique des langues peuvent tirer profit d'informations souvent absentes des dictionnaires construits à la main mais qui peuvent être extraites à partir de corpus (par exemple, la fréquence des entrées ou les têtes lexicales des arguments). Enfin, les moyens techniques (performances des ordinateurs, disponibilité des corpus, outils de traitement automatique de la langue, ...) ont permis d'envisager l'acquisition automatique de lexiques à partir de textes.

Pour toutes ces raisons, les chercheurs en TAL ont tenté d'acquérir automatiquement des informations lexicales à partir de corpus bruts depuis le début des années 1990. Depuis, de nombreuses expériences d'acquisition automatique à partir de corpus brut ont été réalisées sur de nombreuses langues (anglais, allemand, tchèque, grec, italien, français, ...). Ce chapitre présente les principales expériences réalisées. Pour chacun de ces travaux, nous donnons, lorsque cela est possible, un tableau récapitulatif des résultats obtenus. Toutefois, ces résultats sont à considérer avec prudence : ils peuvent varier considérablement selon la granularité des informations obtenues ou la langue observée.

Après avoir donné un aperçu des mesures existantes de la performance de ces systèmes, nous présentons un schéma global commun à la plupart des

méthodes d'acquisition automatique de SSC à partir de corpus. Nous faisons ensuite un état de l'art des premiers travaux réalisés dans le domaine au début des années 90, des travaux réalisés ensuite sur une plus large échelle et des travaux d'acquisition pour le français. Le chapitre se conclut par une synthèse des travaux présentés.

4.1 Méthodes d'évaluation de ressources lexicales

L'acquisition de ressource à partir de corpus est évidemment indissociable des mesures de performance, dans la mesure où les méthodes utilisées n'acquièrent pas des ressources parfaites. L'évaluation doit permettre de valider la qualité intrinsèque des systèmes et leur utilité pour une tâche donnée. Dans la plupart des cas, on évalue les méthodes d'acquisition en mesurant la qualité des informations qu'elles acquièrent (c'est-à-dire des ressources obtenues par l'utilisation de ces techniques). Les ressources lexicales peuvent être évaluées de plusieurs manières. Nous distinguons tout d'abord l'évaluation intrinsèque de l'évaluation extrinsèque puis l'évaluation quantitative de l'évaluation qualitative. La section se termine par une discussion sur les limites des différentes méthodes décrites.

4.1.1 Évaluation intrinsèque et évaluation extrinsèque

En ce qui concerne l'évaluation de ressources, Karen Sparck-Jones et Julia R. Gallier distinguent l'évaluation intrinsèque de l'évaluation extrinsèque (Sparck-Jones et Galliers, 1996). La première consiste à évaluer la ressource en elle même tandis que la deuxième consiste à intégrer cette ressource dans une application pratique et à mesurer son apport à cette tâche.

Évaluation intrinsèque

L'évaluation intrinsèque d'une ressource cherche à montrer si une ressource est correcte « en elle-même ». Pour ce faire, elle examine les informations contenues dans la ressource et évalue leur qualité, soit en les comparant aux informations contenues dans d'autres ressources, soit par introspection.

On peut distinguer deux types d'évaluation intrinsèque : l'évaluation quantitative, qui mesure la qualité de la ressource à l'aide d'outils mathématiques, et l'évaluation qualitative, qui se consacre à une évaluation plus fine des résultats obtenus. Ces deux types d'évaluation sont détaillées plus loin (sections 4.1.2 et 4.1.3).

Évaluation extrinsèque

L'évaluation extrinsèque d'une ressource consiste à montrer qu'une ressource améliore les performances des applications qui en feront usage.

Pour ce faire, on compare les résultats obtenus pour une tâche donnée avec et sans l'exploitation de la ressource. Par exemple, pour ce qui concerne un lexique de sous-catégorisation, on peut considérer une tâche d'analyse syntaxique et comparer les performances obtenues pour un même système dans le cas où il n'utilise pas de ressource et dans le cas où il utilise le lexique que nous cherchons à évaluer (on peut aussi comparer les performances lorsque le système utilise une autre ressource).

Ce type d'évaluation n'est pas facile à mettre en oeuvre. Il suppose en effet de disposer d'un système susceptible d'exploiter les informations disponibles dans la ressource évaluée. En outre, il suppose que l'on ait un moyen fiable d'évaluer les performances de ce système (ce qui peut reposer la question du *gold standard* (cf section 4.1.2)).

Il est parfois plus important de mesurer l'apport d'une ressource par rapport à une tâche donnée, en particulier lorsque cette ressource est destinée à n'être utilisée que pour une tâche précise.

4.1.2 Évaluation intrinsèque : évaluation quantitative

Pour l'évaluation quantitative des ressources, on utilise un *gold standard*, c'est-à-dire une ressource de référence. Cette référence peut être une ressource existante (pour que cette ressource puisse être considérée comme une référence, il faut que sa qualité soit assurée par un certain travail « manuel » : constitution de la ressource ou *a minima*, sa validation), la fusion de plusieurs ressources ou encore elle peut être obtenue à partir de l'annotation manuelle d'un corpus (ce qui permet, en utilisant un extrait du corpus d'acquisition, d'évaluer les performances du système par rapport aux informations contenues dans le corpus).

Les principales mesures d'évaluation utilisées sont la précision, le rappel et la F-mesure.

La précision mesure la proportion d'entrées (c'est-à-dire, pour notre tâche, de couples verbe-SSC) correctes acquises par le système. Elle est mesurée en faisant le rapport entre le nombre de vrais positifs (c'est-à-dire d'entrées présentes à la fois dans la ressource évaluée et dans la référence) et la somme entre le nombre de vrais positifs et le nombre de faux positifs (c'est-à-dire d'entrées présentes dans la ressource évaluée mais pas dans la référence) :

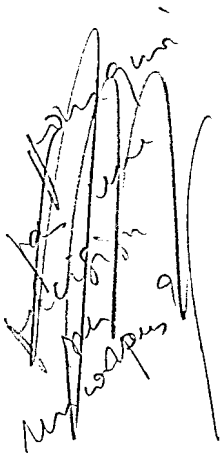
$$Precision = \frac{VP}{VP+FP}$$

Le rappel mesure la proportion d'entrées retrouvées par le système. Il est mesuré en faisant le rapport entre le nombre de vrais positifs et la somme entre le nombre de vrais positifs et le nombre de vrais négatifs (c'est-à-dire d'entrées présentes dans la référence mais pas dans la ressource évaluée) :

$$Rappel = \frac{VP}{VP+VN}$$

La F-mesure est une mesure de la qualité globale du système, c'est-à-dire de l'équilibre entre la précision et le rappel. Elle est mesurée par la formule suivante :

$$F = \frac{2*Precision*Rappel}{Precision+Rappel}$$



Dans certains travaux, une autre mesure de rappel est calculée. Le *token recall* est la proportion de vrais positifs parmi tous les exemples d'un corpus annoté à la main. Cette mesure permet d'évaluer la méthode pour elle-même et non la ressource acquise par son utilisation sur un corpus particulier. En effet, lorsque l'on compare une ressource acquise automatiquement à une référence, des entrées peuvent être absentes de la ressource acquise car elles sont absentes (ou peu présentes) dans le corpus d'acquisition. Dans ce cas, ce n'est pas la méthode d'acquisition qui produit des « erreurs ». Le score de rappel est diminué par de tels cas alors que le *token recall* permet d'éviter ce biais.

Ces quatre mesures sont utilisées pour mesurer les performances de tous les travaux d'acquisition automatique de ressources lexicales. Pourtant, elles ne sont pas toujours pertinentes pour évaluer des systèmes d'acquisition qui n'ont pas l'ambition de concurrencer les dictionnaires acquis manuellement.

En outre, nous devons nous interroger sur le statut du *gold standard* avant d'aller plus loin. En effet, il est difficile d'obtenir une ressource référence qui soit parfaitement adaptée à l'évaluation.

Premièrement, les ressources construites manuellement ne sont pas toujours adaptées à un traitement automatisé (par exemple, même le Lexique-Grammaire, destiné à être utilisé de manière automatique, fait l'objet de tentatives de normalisation pour le rendre plus utilisable par les systèmes de traitement automatique (Gardent et al., 2006; Constant et Tolone, 2010)). Deuxièmement, ces ressources requièrent parfois des transformations de formats coûteuses et/ou susceptibles d'induire des erreurs (c'est notamment le cas de DicoValence). Troisièmement, certaines ressources ne sont pas totalement disponibles. Enfin, les informations disponibles dans ces ressources ne sont pas toujours celles que l'on souhaite évaluer dans les ressources acquises automatiquement (par exemple, les dictionnaires électroniques disposent très rarement d'informations sur la fréquence des entrées).

Le meilleur moyen de constituer une ressource référence est alors de fusionner différents lexiques. Néanmoins, l'hétérogénéité des théories et des formats qui sous-tendent les lexiques construits manuellement rend cette tâche de normalisation ardue et coûteuse.

Il faut donc considérer avec prudence l'évaluation quantitative des ressources par le calcul de la précision, du rappel et de la F-mesure et s'intéresser également à une évaluation plus « qualitative » des lexiques.

4.1.3 Évaluation intrinsèque : évaluation qualitative

Pour évaluer la qualité d'une ressource, on peut également demander à des évaluateurs humains d'examiner ses entrées et de vérifier qu'elles correspondent bien à des schémas de sous-catégorisation corrects. Pour faciliter la tâche des évaluateurs, il est possible d'utiliser des lexiques existants ou d'observer les usages en corpus afin de vérifier s'ils correspondent bien à l'entrée concernée.

L'évaluation qualitative permet d'avoir une vision plus précise des erreurs contenues dans la ressource et de leur origine (par exemple, dans le cas de l'acquisition automatique de schémas de sous-catégorisation : analyse syntaxique erronée, présence de modifieurs, erreur lors de l'acquisition, ...).

Le principal inconvénient de ce type d'évaluation est son coût, ce qui a pour conséquence la grande difficulté à traiter « en largeur » les entrées du lexique par cette approche.

4.1.4 Discussion

Nous venons de voir les principales techniques d'évaluation des systèmes d'acquisition automatique de ressources lexicales. Nous avons montré plus en détail les avantages et les lacunes de ces méthodes d'évaluation dans notre article présenté à la conférence *LREC* en 2008 (Poibeau et Messiant, 2008).

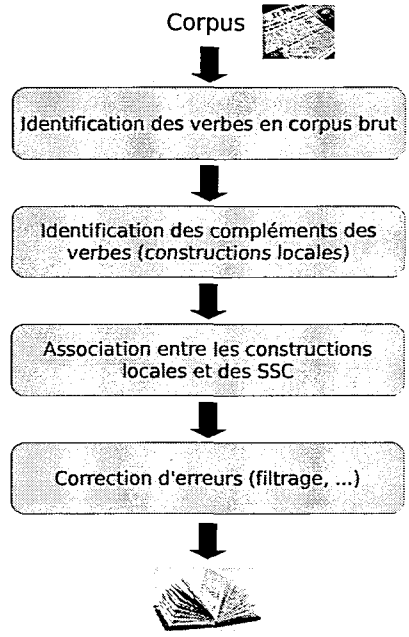


FIGURE 4.1 – *Schéma global des méthodes d'acquisition automatique de schémas de sous-catégorisation*

Étant donné les problèmes liés aux différents types d'évaluation, il nous semble important de mixer plusieurs types d'évaluation. On peut par exemple combiner une évaluation quantitative « en largeur » (à condition de disposer d'un *gold standard* d'une certaine qualité) et une évaluation qualitative, qui permet de juger plus « en profondeur » la qualité de la ressource.

4.2 SCHÉMA GLOBAL POUR L'ACQUISITION AUTOMATIQUE DE SCHÉMAS DE SOUS-CATÉGORISATION

La plupart des travaux d'acquisition automatique de schémas de sous-catégorisation suivent une même approche. En effet, ce processus d'acquisition nécessite des opérations identifiées et les différentes méthodes se distinguent selon l'importance donnée à telle ou telle étape ou selon la façon dont ces opérations sont effectuées. La présentation de ce schéma global permettra de nous concentrer sur les différences entre les travaux présentés ensuite. Le schéma 4.1 donne un aperçu du modèle global décrit dans cette section.

4.2.1 Étape 1 : identification des verbes en corpus

L'acquisition de SSC verbaux nécessite tout d'abord d'identifier les verbes du corpus (avant le développement de lemmatiseurs efficaces, cette étape n'était pas triviale et les stratégies de repérage utilisées pouvaient varier).

Par exemple, dans la phrase suivante, le système d'acquisition doit repérer les verbes *affirmer* et *acheter* :

*Tout d'abord, j'**affirme** que la De Beers n'**a** jamais **acheté** de diamants à l'Unita.*

Deux verbes ont été repérés dans cette phrase : *affirmer* et *acheter*. On remarque déjà une première difficulté pour un système automatique : repérer le verbe *acheter* alors que le verbe est conjugué au passé composé et qu'une forme négative provoque l'insertion du mot *jamais* entre les deux composantes de la forme verbale *a acheté*.

Une fois les verbes identifiés, le système peut passer à la deuxième étape.

4.2.2 Étape 2 : identification des compléments

Il faut ensuite identifier localement les compléments du verbe, c'est-à-dire réaliser une analyse syntaxique de surface au moins partielle. La méthode et les informations retenues (fonctions syntaxiques, catégories morpho-syntaxiques, ...) dépendent de l'approche et/ou de l'analyseur.

Annotons désormais les compléments d'*affirmer* entre crochets et ceux d'*acheter* entre parenthèses dans l'exemple précédent :

*Tout d'abord, [j]'**affirme** [que (la De Beers) n'**a** jamais **acheté** (de diamants) (à l'Unita)].*

Le verbe *affirmer* a donc deux compléments : le pronom personnel *je* (élidé en *j'*) et *que la De Beers n'a jamais acheté de diamants à l'Unita* tandis que *acheter* prend trois compléments : *la De Beers*, *de diamants* et *à l'Unita*. L'imbrication des éléments rend difficile l'analyse.

Maintenant que les verbes et leurs compléments sont identifiés, il faut associer cette configuration locale à un SSC.

4.2.3 Étape 3 : association entre ces constructions locales et des schémas de sous-catégorisation

Le verbe et ses compléments forment une structure argumentale qu'il faut associer à un SSC.

Deux méthodes sont alors possibles :

- Faire correspondre la configuration locale à un SSC dans une liste de schémas existants fournie au système ;
- Inférer un SSC à partir de la configuration locale.

La première méthode permet de se limiter à une liste de schémas en se fondant sur les lexiques existants ou des théories syntaxiques. Elle a l'avantage de mieux traiter les ambiguïtés ou les exceptions pour réduire le nombre de schémas erronés acquis par le système.

La deuxième méthode a l'avantage d'utiliser uniquement les informations disponibles dans le corpus et peut faire émerger des schémas nouveaux par rapport aux lexiques existants qui peuvent être incomplets. Elle n'introduit pas de biais lié à des choix effectués par les linguistes qui auraient constitué les lexiques dont serait issue la liste de SSC : les schémas qui apparaissent dans la ressource sont les schémas présents dans le corpus.

À l'issue de cette étape, le système a associé des SSC aux configurations identifiées plus haut pour notre exemple :

(1) *Tout d'abord, [j]'affirme [que ...].* => `[SUJ:SN, OBJ:COMPL]`

(2) *(la De Beers) n'a jamais acheté (de diamants) (à l'Unita).* => `[SUJ:SN, OBJ:SN, OBJ:SP(à)]`

Le système associe donc une construction complétive `[SUJ:SN, OBJ:COMPL]` à *affirmer* et une construction ditransitive `[SUJ:SN, OBJ:SN, OBJ:SP(à)]` à *acheter* (soit en utilisant une liste de schémas prédéfinie soit en inférant ces schémas à partir du corpus).

4.2.4 Étape 4 : correction d'erreurs

Certains des schémas acquis lors de l'étape précédente sont incorrects et une correction des erreurs est nécessaire pour obtenir la meilleure ressource possible. Il s'agit le plus souvent d'un filtrage des entrées incorrectes (c'est-à-dire d'une sélection parmi les hypothèses produites par l'étape 3) par des méthodes statistiques mais d'autres solutions sont envisageables.

Nous avons présenté un modèle pour l'acquisition automatique de SSC à partir de corpus. Il s'agit d'un schéma global qui permet d'identifier les étapes nécessaires du processus d'acquisition mais les systèmes existants ne suivent pas tous ce modèle « à la lettre » (par exemple, certaines étapes sont parfois omises). Nous allons désormais développer un peu plus les méthodes utilisées pour chacune des étapes nécessaires en parcourant chronologiquement le champ de l'acquisition automatique de SSC à partir de corpus.

4.3 Les premiers travaux d'acquisition automatique de SSC

Les analyseurs syntaxiques statistiques étaient initialement (au début des années 1990) non lexicalisés, ce qui nuisait à leurs performances. Les concepteurs d'analyseurs ont alors remarqué que le repérage de régularités au niveau des corpus permettait d'acquérir automatiquement des informations de type lexical à moindre coût, ce qui équivaut à un apprentissage endogène.

Depuis, avec l'amélioration des techniques statistiques et la plus grande disponibilité de ressources lexicales, ceci est devenu moins nécessaire mais la recherche a continué dans ce domaine afin de permettre l'acquisition de SSC rares ou spécifiques à un domaine. C'est dans cette perspective que nous nous situons.

Si les premiers travaux portaient sur un nombre réduit de verbes et de schémas, la disponibilité d'outils et/ou de corpus a permis d'augmenter la couverture de ces systèmes progressivement pour tendre vers l'exhaustivité.

`NP only / SN seul`	*greet (them).*
`tensed clause / PropSub`	*hope (he'll attend).*
`infinitive`	*hope (to attend).*
`NP & clause / SN et PropSub`	*tell (him) (he's a fool).*
`NP & infinitive / SN et SINF`	*want (him) (to attend).*
`NP & NP / SN et SN`	*tell (him) (the story).*

TABLE 4.1 – *Liste des SSC reconnus par le système Lerner (cette liste est issue de (Brent, 1993))*

4.3.1 Jeter les bases de l'acquisition automatique de SSC à partir de corpus

La première étude d'acquisition automatique de SSC à partir de corpus dont il est fait état dans la littérature est celle de Michael R. Brent (Brent, 1991; 1993). Le but de cette étude est d'explorer l'acquisition d'informations lexicales sur la syntaxe des verbes à partir d'un corpus brut et d'une analyse de surface très partielle.

Le système développé par Brent (appellé **Lerner**) détecte six SSC simples à partir d'un corpus de 2,6 millions de mots (voir tableau 4.1).

Pour reconnaître ces différents schémas dans le corpus, **Lerner** repose sur des indices morpho-syntaxiques locaux plutôt que sur une analyse syntaxique complète. En effet, les applications d'analyse syntaxique étaient à l'époque peu efficaces et coûteuses en ressources. Le système de Brent repose sur l'hypothèse qu'une analyse complète n'est pas nécessaire pour reconnaître les configurations qui correspondent à des SSC.

Cette première étude esquisse déjà les étapes que nous avons mises en évidence dans la section précédente :

- Dans un premier temps, les mots susceptibles d'être des verbes sont repérés dans le corpus (principalement en se basant sur le fait que le mot apparaît avec et sans le suffixe *-ing*).
- La détection des SSC utilise une petite grammaire à états finis qui se concentre sur un fragment de la langue anglaise. Cette grammaire distingue trois types de compléments : les objets directs, les infinitifs et les propositions subordonnées en se basant essentiellement sur des informations de surface du type : « une séquence qui débute par *that the* indique le début d'une proposition subordonnée » ou « si le mot qui suit un verbe est un nom et qu'il n'est pas sujet d'un autre verbe, il a de fortes chances d'être l'objet du verbe ». Dans la phrase *I want to tell him that the idea won't fly.*, **Lerner** utilise le fait que le pronom *him* ne prend presque jamais de proposition relative pour déterminer que *that the* est probablement complément du verbe. De plus, une règle permet au système de reconnaître ce complément comme étant une proposition subordonnée (car il commence par *that the*). C'est donc le schéma `NP & clause / SN et proposition subordonnée` qui est attribué à la phrase.
- La méthode de sélection d'hypothèses utilisée est un filtrage basé sur le test binomial (Kalbfleisch, 1985). Ce test permet de déterminer si un verbe apparaît suffisamment fréquemment avec un SSC pour qu'on associe ce SSC à ce verbe.

En comparant les SSC acquis par leur système avec des schémas associés manuellement à 33 verbes choisis au hasard parmi les 300 verbes les plus courants du corpus, les auteurs obtiennent une précision globale de 86% (83% si on exclut les schémas `REST` (qui correspondent à toutes les constructions qui ne font pas partie des 6 schémas reconnus par le système)).

Cette première étude a permis de poser des bases pour l'acquisition automatique de SSC. Elle a aussi mis en exergue les difficultés inhérentes à la tâche comme la difficulté à traiter des erreurs et à trouver un équilibre entre la précision et le rappel. Dès les débuts de l'acquisition automatique de SSC, on remarque que les modèles statistiques utilisés ne suffisent pas à traiter des erreurs dont la typologie est très variée. Ces difficultés conduisent à faire des choix sur les informations que le système cherche à détecter et à ignorer certains phénomènes de la langue pour obtenir une précision et un temps de calcul satisfaisants plutôt que d'augmenter la couverture en faisant chuter la précision. Par exemple, Brent fait le choix de privilégier la précision à la couverture et son système reconnaît seulement six SSC, ignorant les compléments prépositionnels dont l'ambiguïté pose de nombreux problèmes (tant du point de vue de la distinction argument/modifieur que du point de vue de leur rattachement au bon élément dans la phrase). L'approche de Brent qui repose sur des régularités grammaticales simples n'exploite pas une grande quantité d'informations contenues dans le corpus.

Toutefois, ces limites s'expliquent par la nature exploratoire de cette étude. Ce travail pionnier constitue le socle sur lequel vont se baser la plupart des travaux d'acquisition automatique de SSC à partir de corpus brut.

4.3.2 Mieux exploiter les corpus en les étiquetant *a priori*

Suite aux travaux de Brent, d'autres chercheurs (Ushioda et al., 1993; Manning, 1993) ont tenté d'acquérir des SSC à partir de corpus. Dans ces travaux, les données d'entrée sont découpées en syntagmes par un analyseur partiel à états finis en amont de l'acquisition à proprement parler. Ce type d'analyse facilite le repérage du verbe et de ses compléments.

Notons que l'analyseur syntaxique utilisé pour ce type de travaux doit être non lexicalisé ou qu'il ne doit en tous cas pas disposer d'informations de type « SSC ». En effet, l'acquisition de SSC sur les sorties d'un analyseur lexicalisé ne consisterait qu'à acquérir des informations en grande partie déjà disponibles dans l'analyseur.

Akira Ushioda et al. (1993) utilisent un corpus journalistique (composé d'articles du *Wall Street Journal*) lexicalement étiqueté et un analyseur de syntagmes nominaux à états finis pour reconnaître les six SSC reconnus par Brent (voir tableau 4.1). Des régles d'extraction basées sur des expressions régulières sont ensuite utilisées pour générer des hypothèses. Un jeu de patrons d'erreurs extrait d'un corpus d'entrainement par apprentissage permet de reconnaître les erreurs types pour filtrer ces hypothèses. L'apport principal de cette étude est l'acquisition des fréquences relatives des SSC pour un verbe donné en corpus, qui sont obtenues en faisant le rapport entre le nombre d'occurences de la combinaison SSC-verbe et le nombre d'occurences du verbe. Ceci est utile car une telle information est exploitée par

certaines applications de TAL (par exemple, des analyseurs syntaxiques probabilistes).

L'algorithme d'acquisition de (Manning, 1993) prend en entrée un corpus de 4 millions de mots traité par un étiqueteur de parties du discours. Dans un premier temps, un analyseur syntaxique à états finis basé sur l'ordre des mots en anglais parcourt ce corpus. Il détecte les compléments des verbes ainsi que des statistiques sur l'apparition des verbes dans divers contextes. Ces statistiques permettent ensuite de filtrer les sorties de l'analyseur par un test binomial et d'associer les verbes à des SSC (parmi une liste prédéfinie de 19 schémas). Cette méthode a permis d'acquérir un dictionnaire de près de 5000 entrées pour plus de 3000 verbes mais qui ne comprend pas d'informations sur la fréquence des arguments dans le corpus. Contrairement aux travaux présentés précédemment, cette méthode cherche à acquérir des SSC contenant des compléments prépositionnels. Manning est donc confronté au problème de la distinction argument/modifieur. Son analyseur de surface ne faisant pas de distinction entre arguments et modifieurs, la solution choisie est de chercher à éliminer les schémas lors de l'étape de filtrage. Il est en effet difficile d'intégrer les critères définis par les linguistes au sein de systèmes de traitement automatique et on peut faire l'hypothèse que les schémas contenant des modifieurs seront moins fréquents en corpus que ceux qui n'en contiennent pas (en raison du caractère optionnel des modifieurs, voir section 2.3.2). Depuis plus de quinze ans, la solution choisie par Manning reste la principale réponse apportée au problème de la distinction argument/modifieur dans les systèmes d'acquisition automatique.

Pour évaluer ce système, Manning a sélectionné au hasard 40 verbes parmi 2 000 verbes courants. La comparaison des SSC acquis avec les schémas présents dans le dictionnaire OALD donne un taux de précision de 90% et un rappel de 43%.

Les apports des approches présentées ci-dessus sont considérables par rapport à la méthode de Brent. L'utilisation d'une étape de segmentation en syntagmes (*chunking*) permet de mieux exploiter les informations contenues dans le corpus en produisant un plus grand nombre d'hypothèses en amont de l'étape de filtrage. De plus, ces techniques ont permis de s'attaquer à des problèmes importants en acquisition automatique, comme la distinction argument/modifieur et l'extraction d'informations de fréquence des SSC. Néanmoins, ces méthodes ne permettent pas encore d'acquérir des SSC à une large échelle et se limitent au mieux à quelques dizaines de schémas.

4.4 Acquisition de SSC à large échelle

Le développement d'outils de TAL comme les étiqueteurs morpho-syntaxiques ou les analyseurs syntaxiques plus complets et la disponibilité de corpus de taille plus conséquente ont permis d'envisager l'acquisition de SSC à une plus large échelle. C'est notamment le cas des travaux menés à Cambridge depuis le milieu des années 90 (Briscoe et Carroll, 1997; Korhonen et al., 2000; Korhonen, 2000; 2002; Korhonen et al., 2006; Preiss et al., 2007).

Le système d'acquisition de (Briscoe et Carroll, 1997) réalise les traitements correspondant aux quatre étapes de l'acquisition sur les phrases du corpus d'entrée [1] :

1. Identification des verbes en corpus : Un *étiqueteur* étiquette la phrase en parties du discours et un *lemmatiseur* remplace les paires mot-étiquette par des paires lemme-étiquette en fonction des parties du discours associées aux éléments par l'étiqueteur ;

2. Identification des compléments et des constructions locales : un *analyseur syntaxique probabiliste non lexicalisé* (**RASP (Robust Accurate Statistical Parsing)**), entrainé sur un corpus arboré, produit des analyses syntaxiques (hierarchisées) en constituants pour chaque phrase (Carroll et Briscoe, 1996). Ensuite, un *extracteur de patrons* extrait des patrons de sous-catégorisation (c'est-à-dire des constructions locales), incluant les catégories syntaxiques et les lemmes des têtes des constituants, à partir des analyses syntaxiques de chaque phrase ;

3. Association entre ces patrons de sous-catégorisation locaux et des SSC : un *classifieur de patrons* assigne des SSC aux patrons extraits ou les déclare inclassifiables en se basant sur les caractéristiques des catégories syntaxiques et des lemmes des têtes des constituants. Le système utilise une liste de schémas constituée en combinant des dictionnaires de sous-catégorisation existants pour l'anglais (*ANLT* et *COMLEX*). Nous avons déjà évoqué le fait qu'une telle liste peut augmenter la précision du système mais qu'elle risque de réduire son rappel en ne lui permettant pas d'acquérir des informations présentes dans le corpus mais absentes des dictionnaires utilisés pour constituer la liste ;

4. Filtrage des erreurs : un *évaluateur de patrons* filtre les SSC candidats issus du classifieur à l'aide d'un test binomial. Ce filtre permet d'exclure des SSC incorrects (par exemple présentant des modifieurs).

Ce système permet d'associer aux verbes repérés en corpus 160 SSC différents. Le système extrait également la fréquence relative des entrées. Une première expérience a permis d'évaluer le système sur une liste de 14 verbes en comparant le lexique obtenu avec les SSC issus de dictionnaires existants (Briscoe et Carroll, 1997). Les résulats de cette évaluation font état d'un taux de précision de 65,7% et un taux de rappel de 35,5%. L'évaluation des résultats par rapport à une analyse manuelle du corpus donne des résultats sensiblement meilleurs : 76,6% en précision et 43,4% en rappel. Le *token recall* pour ces sept verbes est de 80,9%. Ces résultats montrent que l'évaluation des SSC acquis par rapport à un dictionnaire peut biaiser les résultats. En effet, la présence dans le corpus de constructions non répertoriées dans le dictionnaire ou l'absence de schémas rares ou anciens dans le corpus ne permet pas de juger de la performance du système « en lui-même ». Nous reviendrons sur ce problème dans le chapitre 6.

Les auteurs ont également montré que les SSC extraits par leur système améliorent sensiblement les performances d'un analyseur syntaxique malgré

1. L'article présente six composants qui sont chargés d'effectuer en série les traitements nécessaires qui correspondent *grosso modo* aux étapes décrites au début de ce chapitre.

les taux d'erreurs cités ci-dessus (Briscoe et Carroll, 1997; Carroll et al., 1998).

Depuis ce premier article, la méthode a fait l'objet de nombreuses améliorations jusqu'à très récemment.

Anna Korhonen et al. (2000) compare trois méthodes de filtrage différentes pour ce système d'acquisition : la loi binomiale, souvent utilisée (*binomial hypothesis testing*) (Brent, 1993; Manning, 1993), la log-vraisemblance (*binomial log-likelihood ratio* (Gorrell, 1999)) et le maximum de vraisemblance (*maximum likelihood estimates*). La technique consiste à appliquer un seuil sur les fréquences relatives des SSC déterminé empiriquement à l'aide d'un jeu de test. La méthode peut être améliorée en utilisant des seuils différenciés en fonction des schémas évalués (par exemple, en réduisant le seuil pour des constructions relativement rares en corpus mais « sûres » ou en l'augmentant pour des constructions fréquentes mais rarement correctes comme celles contenant certains modifieurs). C'est la méthode de maximum de vraisemblance qui donne de meilleurs résultats sur 14 verbes et qui a été retenue par Korhonen. Selon l'auteur, celà s'explique par la distribution zipfienne des SSC et la faible corrélation entre les distributions conditionnelles et inconditionnelles. Néanmoins, le filtrage par un seuil rejette tous les SSC peu fréquents dans le corpus alors que dans certains cas, des schémas rares sont tout à fait pertinents.

Pour remédier au problème du filtrage des SSC rares, Anna Korhonen propose de compenser des informations absentes ou trop peu présentes du corpus à l'aide de classes de verbes (*back-off estimates*) (Korhonen, 2000). En effet, il y a une corrélation entre la présence d'un verbe dans une classe (syntaxique ou sémantique) et ses constructions syntaxiques. Autrement dit, les verbes d'une même classe ont des chances de partager les mêmes SSC. En utilisant les classes de Levin (Levin, 1993) et la classification syntaxique du dictionnaire ANLT, les auteurs proposent de « lisser » les distributions de probabilités des SSC avant l'étape de filtrage. Trois méthodes de « lissage » sont utilisées : *add-one smoothing*, *Katz backing-off* et l'interpolation linéaire. Elles permettent toutes les trois d'augmenter la précision et le rappel des SSC obtenus, l'interpolation linéaire étant la méthode la plus efficace. Le « lissage » des probabilités de distribution est donc efficace pour répondre au problème du manque de données pour certains verbes ou schémas. Notons néanmoins qu'un certain effort manuel est nécessaire pour utiliser cette méthode à large échelle.

En 2006, des techniques pour améliorer les résultats du système ont été proposés (Korhonen et al., 2006). Les auteurs ont dérivé quatre lexiques à partir du lexique non filtré :

1. un lexique filtré par un seuil sur les fréquences relatives des entrées ;
2. un lexique créé en « lissant » les distributions de schéma par interpolation linéaire puis en filtrant le résultat à l'aide d'un seuil ;
3. un lexique créé en sélectionnant les SSC présents à la fois dans le lexique non filtré et dans les dictionnaires ANLT et/ou COMLEX puis en ajoutant au résultat les SSC dont la fréquence relative dépasse un seuil spécifique pour chaque schéma ;
4. un lexique créé en utilisant la méthode précédente (3) mais les SSC

ajoutés sont d'abord « lissés » par interpolation linéaire avant d'être filtrés puis ajoutés au lexique.

Les résultats obtenus *via* la dernière méthode sont très bons (93% de précision et 82% de rappel) mais nécessitent des dictionnaires construits manuellement (cette approche n'est donc pas totalement automatique, ce qui la rend difficilement reproductible (par exemple, sur une langue de spécialité)). Par ailleurs, nous souhaitons quant à nous éviter d'utiliser des ressources existantes, dans la mesure où nous nous plaçons justement dans un contexte d'acquisition.

Enfin, Judita Preiss & al. montrent que ce système d'acquisition peut également être adapté pour l'acquisition de SSC d'adjectifs et de noms. La méthode permet alors de distinguer 168 schémas verbaux, 37 schémas adverbiaux et 31 schémas nominaux (Preiss et al., 2007).

Les travaux présentés jusqu'ici concernent l'anglais mais l'acquisition automatique de SSC a fait l'objet d'études sur d'autres langues. Le schéma global présenté au début de ce chapitre n'est pas lié à une langue particulière et la plupart des travaux présentés ici s'en inspirent largement. D'autres travaux s'inspirent en grande partie du schéma global et des travaux menés à Cambridge et ont adapté ce schéma pour de nombreuses langues : l'allemand (Schulte im Walde, 2002; Schulte im Walde, 2009), l'italien (Lenci et al., 2008), le hongrois (Serény et al., 2008), le grec (Maragoudakis et al., 2001), le turc (caslan et al., 2007), le chinois (Han et al., 2004) ou le bengali (Banerjee et al., 2009) par exemple. Toutefois, les spécificités des langues imposent le développement de nouveaux systèmes (contrairement, par exemple, à certains outils d'apprentissage automatique qui peuvent être utilisés sur différentes langues moyennant quelques légers réglages). Par exemple, un système consacré à l'acquisition en français ne saurait ignorer le problème des constructions pronominales.

4.5 Travaux d'acquisition pour le français

Nous présentons ici deux études concernant l'acquisition automatique de SSC pour le français. À notre connaissance, il n'existe pas d'autres travaux de ce type pour la langue française.

Paula Chesley et Susanne Salmon-Alt présentent une première étude exploratoire d'extraction automatique de SSC pour le français (Chesley et Salmon-Alt, 2006). La méthode d'acquisition employée repose sur le schéma global défini au début de ce chapitre. Le système prend en entrée les sorties d'un analyseur syntaxique chargé de réaliser le repérage des verbes et des compléments. L'acquisition de SSC en elle-même nécessite deux étapes : pré-filtrage et filtrage.

L'étape de pré-filtrage prend en entrée le corpus analysé par l'analyseur syntaxique VISL et extrait les SSC pour la liste des verbes dont on souhaite acquérir les SSC. Le système n'inclut pas de liste de SSC *a priori*, les schémas sont déterminés au cours du processus en fonction de l'analyse syntaxique. Les éléments constituants des SSC sont les objets directs, les syntagmes prépositionnels (les prépositions sont spécifiées), les propositions

subordonnées, les verbes à l'infinitif, les syntagmes adjectivaux et les clitiques réflexifs. Les sujets n'apparaissent pas dans les schémas car les verbes prennent obligatoirement un sujet en français. L'analyseur syntaxique utilisé ne fait pas de distinction entre arguments et modifieurs. Les schémas extraits à la première étape incluent donc arguments et modifieurs. A l'image de (Manning, 1993), les auteurs font l'hypothèse que le filtrage éliminera les SSC incorrects (comprenant des modifieurs). Les sorties de cette première étape sont bruitées et une étape de filtrage est nécessaire. La méthode de filtrage utilisée à la deuxième étape est la distribution binomiale des schémas. Une évaluation sur 104 verbes par deux évaluateurs a donné un score de précision entre 86,8% et 96,4%[2][3]. Le *token-recall* (mesuré en examinant quatre phrases extraites aléatoirement d'un corpus journalistique pour chacun des verbes) est de 54,3%. Si cette première étude a permis de montrer que l'acquisition automatique de SSC à partir de corpus brut est possible pour le français, elle n'a été testée qu'à une échelle limitée (104 verbes, 27 SSC uniques et 176 combinaisons verbe/schéma).

Une autre étude d'acquisition automatique de SSC à partir de corpus a été menée dans le cadre du projet *TALC (Traitement automatique des langues et des connaissances)*[4] (Gardent, 2009; Gardent et Lorenzo, 2010). Claire Gardent a mené une expérience portant sur un nombre beaucoup plus important de verbes en partant d'un corpus de 170 millions de mots[5]. Ce corpus est ensuite analysé au moyen de l'analyseur syntaxique TagParser mis au point par Gil Francopoulo (Francopoulo, 2005). Le repérage de régularités au niveau des compléments du verbe permet d'inférer des SSC pour chaque verbe, suivant une stratégie proche de celle de l'équipe de Cambridge. Le lexique résultant, **EasyLex**, est disponible sur le portail TALC (`http://talc.loria.fr/Easylex.html`). Il concerne plus de 4 800 verbes et fournit une moyenne de 6 schémas par verbe. Le lexique n'est pas téléchargeable librement mais une interface de démonstration lui est consacrée sur le portail du projet TALC : `http://talc.loria.fr/Easylex.html`. EasyLex est au format EASy, ce qui facilite son utilisation dans des applications de traitement automatique des langues.

4.6 Conclusion

Le tableau 4.2 présente une synthèse des travaux d'acquisition automatique de SSC à partir de corpus présentés jusqu'ici (corpus utilisé, nombre de verbes traités et nombre de SSC différents obtenus) tandis que le tableau 4.3 présente les résultats obtenus par ces différents systèmes d'acquisition (nombre de verbes évalués, précision, rappel et F-mesure).

Après avoir fait l'état de l'art des méthodes d'évaluation de ressources lexicales, nous avons présenté dans ce chapitre un schéma global pour l'ac-

2. La borne inférieure correspond aux entrées jugées correctes par les deux évaluateurs et la borne supérieure à celles jugées correctes par au moins l'un des deux évaluateurs.

3. L'écart entre ces deux bornes démontre la difficulté à évaluer ce type d'informations lexicales.

4. `http://talc.loria.fr/`

5. Il s'agit du Corpus Passage Court — CPC `http://atoll.inria.fr/passage/ressources.en.html`.

Système	Corpus	Nb verbes	Nb SSC
Brent (1991, 1993)	Brown corpus	?	6
Ushioda & al (1993)	Wall Street Journal	?	6
Manning (1993)	New York Times	3 104	19
Briscoe & Carroll (1997)	Susanne, SEC et LOB	?	160
Korhonen (2000)	BNC	?	163
Korhonen (2006)	6 ressources	6 397	163
Chesley & Salmon-Alt (2006)	Frantext (littéraire)	104	27
Gardent (2009)	Corpus Passage Court	4 800	?

TABLE 4.2 – *Synthèse des méthodes d'acquisition automatique de SSC présentées (les « ? » correspondent à des données non disponibles ; pour Korhonen (2006), il s'agit des chiffres pour le lexique 3)*

Système	Nb verbes évalués	Précision	Rappel (type recall)	F-Mesure
Brent (1993)	?	0.96	0.60	?
Ushioda & al (1993)	33	0.86	?	?
Manning (1993)	40	0.90	0.43	?
Briscoe & Carroll (1997)	14	0.66	0.35	?
Korhonen (2000)	14	0.75	0.58	0.65
Korhonen (2006)	183	0.76	0.63	0.69
Chesley & Salmon-Alt (2006)	24	0.87	0.54 (token recall)	?

TABLE 4.3 – *Synthèse des méthodes d'acquisition automatique de SSC : évaluation (les « ? » correspondent à des données non disponibles ; pour Korhonen (2006), il s'agit des chiffres pour le lexique 3)*

quisition automatique de SSC à partir de corpus et les principales expériences réalisées pour l'anglais. Nous avons enfin décrit les principaux travaux d'acquisition existants pour le français.

Si les méthodes utilisées pour les premières tentatives d'acquisition à partir de corpus brut étaient très variées, de nombreuses études récentes utilisent le même modèle. En effet, la disponibilité et la qualité des étiqueteurs morpho-syntaxiques et des analyseurs syntaxiques simplifient les premières étapes (identification des verbes, extraction des arguments, ...). Ainsi, les travaux récents se concentrent plutôt sur l'acquisition des schémas à proprement parler et le filtrage des entrées non pertinentes, notamment en raison de la distinction argument/modifieur. Le système d'acquistion automatique de SSC pour les verbes français présenté au chapitre suivant s'inscrit dans la lignée de ces travaux.

5

Description du système d'acquisition de schémas de sous-catégorisation pour le français

Nous l'avons vu dans le chapitre précédent, il existe peu de travaux concernant l'acquisition automatique de SSCs pour le français à partir de corpus. Pourtant, les outils de TAL comme les analyseurs syntaxiques ont de plus en plus besoin de ce type de ressources. En outre, l'étude des SSC obtenus par de telles méthodes et leur comparaison avec les lexiques existants peut être très intéressante mais a été relativement peu abordée dans la littérature.

Nous présentons dans ce chapitre notre système d'acquisition de schémas de sous-catégorisation ASSCi (Acquisition de schémas de sous-catégorisation informatisée). Après une présentation de l'architecture globale d'ASSCi, nous présentons les deux outils utilisés pour effectuer les prétraitements puis les trois modules qui composent notre système. Le chapitre se termine par une discussion sur le système ASSCi.

5.1 Architecture générale d'ASSCi

La plupart des systèmes d'acquisition automatique partagent le même modèle général (voir section 4.2). En effet, quels que soient les choix effectués par les concepteurs, ces systèmes ont tous la même visée et il est possible d'identifier les principales tâches qu'ils doivent effectuer.

L'architecture d'ASSCi repose sur ce modèle global. Nous rappelons ci-dessous les quatre étapes principales de ce modèle :

1. l'identification des verbes en corpus ;
2. l'identification des compléments de ces verbes, c'est-à-dire des configurations syntaxiques locales (ou constructions locales) ;
3. l'association entre ces constructions locales et des schémas de sous-catégorisation ;
4. la correction d'erreurs *via* un filtrage qui permet de distinguer les SSC erronés des SSC corrects pour chaque verbe.

Nous avons montré dans le chapitre 4 qu'il existe des variations quant à la façon dont ces tâches sont réalisées. Pour ce qui concerne ASSCi, l'utilisation d'outils pour le pré-traitement des données permet de faciliter les deux premières tâches. De plus l'étape 3 du schéma global est effectuée sans l'aide d'une liste préétablie de SSC, ce qui permet de ne pas se limiter aux seuls SSC connus et donc d'acquérir les informations contenues dans le corpus sans *a priori*. Nous soulignerons ces spécificités tout au long de la description du système.

La figure 5.1 montre la chaîne de traitement de notre système d'acquisition et ses différents modules. ASSCi est constitué de quatre modules qui permettent d'exécuter en série les tâches décrites ci-dessus :

1. **pré-traitements** : le corpus brut est lemmatisé et annoté par TreeTagger puis analysé par Syntex ;
2. **extracteur de pré-schémas de sous-catégorisation** : pour chaque phrase, on identifie les compléments du/des verbe(s) et on extrait les informations utiles pour constituer les SSC candidats (constructions locales) ;
3. **constructeur de schémas candidats** : on constitue un lexique de « base » en inférant des SSC à partir des pré-schémas observés en corpus pour chaque verbe ;
4. **filtre des SSC non pertinents** : on élimine les schémas incorrects par des méthodes statistiques.

A l'issue du processus d'acquisition, le système produit un lexique composé de couples lemme verbal - SSC ainsi que d'informations statistiques et lexicales associées à ces couples (la composition détaillée des entrées du lexique sera présentée dans la section 6.1). Les spécificités de chacun des modules d'ASSCi sont présentées à la section 5.4. Nous présentons tout d'abord les outils utilisés pour effectuer les pré-traitements nécessaires à l'acquisition.

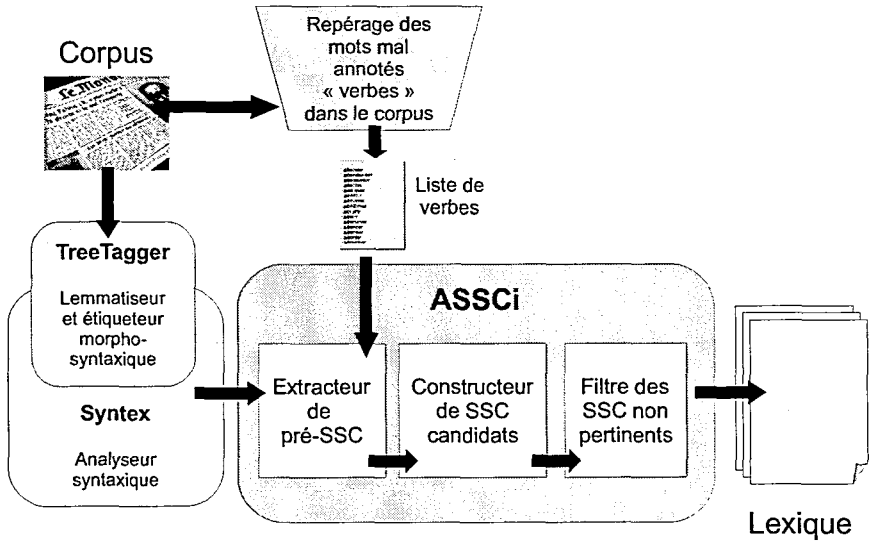


FIGURE 5.1 – *Architecture générale d'ASSCi*

5.2 PRÉ-TRAITEMENTS

Les deux premières étapes de la tâche d'acquisition consistent à identifier les verbes et leurs compléments dans le corpus. Ces tâches peuvent être grandement facilitées par la lemmatisation, l'analyse morpho-syntaxique et l'analyse syntaxique de surface du corpus. Au regard des outils disponibles pour le français et des performances de ceux-ci, nous avons retenu l'analyseur syntaxique SYNTEX. Cet analyseur repose sur les annotations de l'analyseur morpho-syntaxique TREETAGGER. Cette section présente ces deux outils.

5.2.1 Lemmatisation et étiquetage morpho-syntaxique : TreeTagger

La lemmatisation et l'étiquetage morpho-syntaxique sont préalablement réalisés par TREETAGGER (Schmid, 1994) [1].

Description

TREETAGGER est un outil de lemmatisation et d'annotation en parties du discours. Il a été utilisé pour étiqueter des textes en allemand, anglais, français, italien, néerlandais, espagnol, bulgare, russe, grec, portugais et chinois. De plus, il est adaptable à d'autres langues moyennant la disponibilité d'un lexique et d'un corpus annoté.

Dans un premier temps, TREETAGGER segmente le texte en phrases et les phrases en mots [2]. Ensuite, l'étiquetage morpho-syntaxique permet d'associer une étiquette morpho-syntaxique à chaque mot : TREETAGGER estime

1. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
Les fichiers de configuration de TREETAGGER pour le français ont été développés par Achim Stein et Michel Généreux.

2. Le fichier de paramètrage du segmenteur pour le français a été fourni

la probabilité pour chaque mot de porter différentes étiquettes possibles en utilisant des lexiques et le contexte du mot, c'est-à-dire les mots proches de lui dans le corpus[3]. Enfin, un lemme est associé à chaque mot du texte (lemmatisation).

Exemple

Tout au long de ce chapitre, nous allons suivre le processus d'acquisition à l'aide d'un exemple.

Considérons la phrase :

> *Les fameuses « erreurs » de Sartre, il les lui reproche autant que ses adversaires, mais il les lui reproche au nom du Sartre qu'il aime, et qui lui ressemble, l' homme seul, le Roquentin.*

Pour cette phrase, TreeTagger produit l'annotation représentée dans le tableau 5.1. La liste complète des étiquettes utilisées par TreeTagger pour l'annotation du français est disponible sur le site de l'étiqueteur[4].

Chaque ligne représente l'étiquetage correspondant à un mot de la phrase. La colonne de gauche fournit les étiquettes morpho-syntaxiques produites par TreeTagger tandis que la colonne de gauche fournit les lemmes correspondant aux éléments. Par exemple, *aime* est annoté `VER:pres aimer` : il s'agit d'un verbe (`VER`) au présent (`pres`) dont le lemme est *aimer* et *il* est un pronom personnel dont le lemme est *il* (`PRO:PER il`).

Discussion

L'un des avantages de TreeTagger, outre sa robustesse et son efficacité, est son ouverture : il est possible d'ajouter des traitements en amont (par exemple, faire à sa place le découpage en unités de traitement (*tokenisation*)). Ainsi, des règles et des lexiques de reconnaissance des unités syntaxiques complexes (e.g. locutions prépositionnelles) ont été ajoutés pour Syntex (Bourigault et al., 2005). Il est également possible d'intégrer dans la chaîne de traitement des règles de *tokenisation* et de pré-étiquetage spécifiques au corpus à analyser, ce qui est fondamental lorsque l'étiqueteur doit traiter des données non standard (codes de produits, nomenclature d'éléments chimiques, *etc.*). L'analyseur syntaxique a enfin la possibilité de faire des retours en arrière sur l'étiquetage et de modifier les étiquettes attribuées par TreeTagger.

Pour toutes ces raisons, TreeTagger est un outil pertinent pour l'étiquetage morphosyntaxique du français dans le cadre de notre travail.

---

par Michel Généreux : `ftp://ftp.ims.uni-stuttgart.de/pub/corpora/french-chunker-par-linux-3.1.bin.gz`.

3. Les fichiers de paramètrage de TreeTagger en français sont fournis par Achim Stein : `http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html`.

4. `http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html`

DET : ART	le
ADJ	fameux
NOM	«
NOM	erreur
VER : pper	»
PRP	de
NOM	Sartre
PUN	,
PRO : PER	il
PRO : PER	la|le
PRO : PER	lui
VER : subp	reprocher
ADV	autant
KON	que
DET : POS	son
NOM	adversaire
PUN	,
KON	mais
PRO : PER	il
PRO : PER	la|le
PRO : PER	lui
VER : subp	reprocher
PRP : det	au
NOM	nom
PRP : det	du
NOM	Sartre
PRO : REL	que
PRO : PER	il
VER : pres	aimer
PUN	,
KON	et
PRO : REL	qui
PRO : PER	lui
VER : pres	ressembler
PUN	,
DET : ART	le
NOM	homme
ADJ	seul
PUN	,
DET : ART	le
NOM	roquentin
SENT	.

TABLE 5.1 – *Exemple d'annotation par* TREETAGGER *: la colonne de gauche fournit l'étiquette morphosyntaxique tandis que la colonne de droite est le lemme du mot (la liste des étiquettes utilisées par* TREETAGGER *pour le français est donnée en annexe A.*

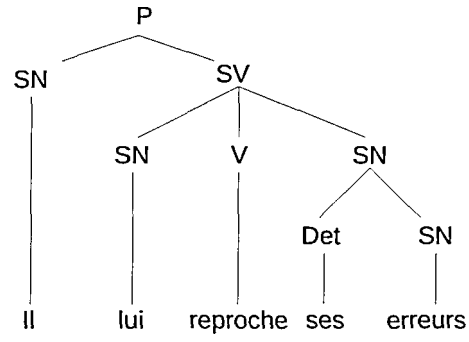


FIGURE 5.2 – *Analyse syntaxique en constituants*

5.2.2 Analyse syntaxique de surface : Syntex

Le corpus d'acquisition est ensuite analysé par SYNTEX, analyseur syntaxique en dépendances développé par Didier Bourigault[5] (Bourigault et al., 2005; Bourigault, 2007).

Description

SYNTEX est un analyseur syntaxique en dépendances dont l'objectif est de traiter des corpus de grande taille composés de phrases du français courant. L'analyse en dépendances s'oppose à l'analyse en constituants : un analyseur en constituants constitue des groupes correspondant à des unités syntaxiques (groupe verbal, ...) tandis qu'un analyseur en dépendances relie les différents mots de la phrase entre eux à partir d'un élément racine (en général, le verbe). Les schémas 5.2 et 5.3 représentent respectivement les analyses en constituants et en dépendances de la phrase *Il lui reproche ses erreurs.*. Les stemmas de Tesnière (Tesnière, 1959) sont une représentation possible de l'analyse en dépendances (voir la figure 2.1 dans la section 2.2 pour un exemple de stemma). L'analyse en dépendances a l'avantage d'être plus indépendante des théories linguistiques que l'analyse en constituants. Pour décrire les relations entre les mots dans une analyse en dépendances, on utilise les concepts de recteur (ou gouverneur) et de régi (ou de dépendant). Par exemple, dans l'arbre 5.3, *reproche* est le recteur de *il*, *lui* et *erreurs* et *ses* est régi par *erreurs*.

Il y a une équivalence « faible » entre ces deux types d'analyse. On peut en général passer de l'analyse en dépendances à l'analyse en constituants (on peut retrouver les groupes en suivant les relations recteurs / régis) alors que la réciproque est moins évidente (on peut difficilement retrouver les relations recteurs / régis en ne disposant que des constituants).

Les principales relations syntaxiques reconnues par SYNTEX sont les suivantes : sujet, complément d'objet direct, complément prépositionnel (de

5. Le développement de SYNTEX fait suite à celui de LEXTER, analyseur dédié au repérage des syntagmes nominaux dans les corpus spécialisés.

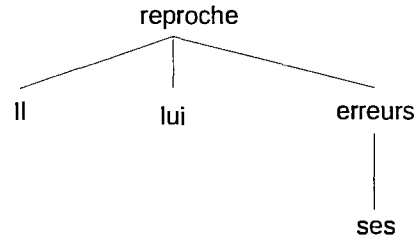


FIGURE 5.3 – *Analyse syntaxique en dépendances*

nom, de verbe et d'adjectif), antécédence relative (*i.e.* antécédent des pronoms relatifs), modification adjectivale (épithète, attribut) et subordination. Chaque élément de la phrase est annoté par ses relations de recteur ou de régi avec les autres éléments.

Pour annoter les éléments de la phrase, SYNTEX applique différents modules de reconnaissance de relations syntaxiques en série : chaque module prend en charge une relation syntaxique particulière et l'entrée de chaque module est la sortie du module qui le précède. Cependant, des retours en arrière sont possibles dans la chaîne de traitement et un module peut remplacer l'étiquette apposée par un module qui l'a précédé, si nécessaire. Ce fonctionnement rend malgré tout essentiel le choix de l'ordre d'exécution des modules.

SYNTEX repose fondamentalement sur des règles et des procédures d'apprentissage endogène, même si des informations lexicales ponctuelles sont utilisées en fonction des besoins.

En effet, SYNTEX n'utilise pas de lexique de sous-catégorisation : il dispose simplement de quelques listes, par exemple des listes de couples verbe / préposition (auxquels sont associés des probabilités acquises à partir d'un grand corpus) pour faciliter le traitement des locutions prépositionnelles. Les interconnexions qui peuvent exister entre divers arguments de la structure argumentale (par exemple, le fait que le verbe n'accepte pas telle préposition en l'absence d'un complément d'objet nominal) ne sont donc pas prises en compte par SYNTEX, ce qui justifie l'utilisation de cet analyseur pour l'acquisition de SSC[6].

Arguments et modifieurs sont rattachés au verbe sans qu'il soit fait de distinction entre eux. Une meilleure prise en compte de la sous-catégorisation par SYNTEX permettrait d'améliorer les performances de l'analyseur en considérant l'intégralité de la structure argumentale et plus seulement des relations locales.

Exemple

Par exemple, pour la séquence « *Les fameuses « erreurs » de Sartre, il les lui reproche autant que ses adversaires, mais il les lui reproche au nom du Sartre*

6. Nous l'avons évoqué au chapitre précédent, l'acquisition de SSC à partir d'un analyseur « très » lexicalisé pose la question de la différenciation entre ce qui est acquis par le système et les lexiques dont dispose l'analyseur syntaxique.

qu'il aime, et qui lui ressemble, l' homme seul, le Roquentin. », SYNTEX produit l'analyse fournie dans la figure 5.4[7]. La figure 5.5 donne une représentation de l'annotation de la séquence « *il les lui reproche au nom du Sartre qu'il aime, et qui lui ressemble, l' homme seul, le Roquentin.* » par SYNTEX.

Chaque ligne contient les informations relatives à un élément de la phrase. La première colonne représente le rang (identifiant) de l'élément. La deuxième colonne est le lemme tandis que la troisième colonne contient la catégorie morpho-syntaxique issue de l'annotation par TREETAGGER. Les deux dernières colonnes sont le(s) recteur(s)[8] et le(s) régi(s) de l'élément[9] (pour chaque recteur et régi, SYNTEX fournit sa relation à l'élément et son rang).

Par exemple, l'analyse produite pour la deuxième occurrence du verbe *reprocher* :

```
21 | reprocher | VCONJS | | SUJ;18, OBJ;19, PREP;20, PREP;22.
```

L'élément de rang 21 de la phrase est le verbe (catégorie VCONJS, verbe conjugué singulier) *reproche* dont le lemme est *reprocher*. Cet élément n'a pas de recteur. Les régis de *reprocher* sont l'élément de rang 18 (sujet *il*), l'objet de rang 19 (*le*) et les éléments 20 (complément prépositionnel *lui*) et 22 (complément prépositionnel *au nom de*).

Discussion

Le choix de SYNTEX comme analyseur syntaxique pour notre système d'acquisition automatique de SSC s'explique à la fois par ses propriétés, ses performances sur les corpus écrits et par sa disponibilité. Les résultats de SYNTEX lors de la campagne d'évaluation EASY (Evaluation des Analyseurs SYntaxiques) en 2007[10] montrent que l'analyseur obtint alors les meilleurs résultats en précision et en F-mesure sur les corpus écrits[11]. De plus, la robustesse de SYNTEX garantit son adaptabilité, du moins sur les textes qui respectent les normes syntaxiques de l'écrit standard. La réutilisation de notre système d'acquisition sur des corpus de domaines particuliers (médecine, droit, ...) ne devrait donc pas poser de problème.

Néanmoins, il existe encore une part non négligeable d'erreurs dues à des faiblesses de l'étiquetage morpho-syntaxique ou de l'annotation des relations, ce qui est inévitable étant donné l'état de l'art. La plupart de ces erreurs se répercuteront dans le processus d'acquisition des SSC et devront être traitées ou filtrées au cours du processus. Toutefois, les erreurs de TREETAGGER qui concernent le repérage des verbes doivent être traitées *a priori*. Le script présenté à la section suivante a pour objectif de répondre à ce problème.

7. Il s'agit d'une représentation de l'analyse syntaxique. Pour un aperçu du format de SYNTEX, voir l'annexe B.

8. En général, les éléments n'ont qu'un seul recteur.

9. SYNTEX fournit également la forme fléchie de l'élément.

10. http://www.limsi.fr/Recherche/CORVAL/easy/

11. L'ensemble des résultats de SYNTEX à la campagne EASY sont disponibles sur la page consacrée au logiciel : http://w3.erss.univ-tlse2.fr/membres/bourigault/syntex.html. On remarquera le différentiel de performance important entre SYNTEX et ses poursuivants en terme de précision sur les corpus écrits lors de cette campagne d'évaluation.

1	le	Det ? ?	DET ;4	
2	fameux	AdjFP	ADJ ;4	
3	«	Typo		
4	erreur	Nom ?P		DET ;1, ADJ ;2, PREP ;6
5	»	Typo		
6	de	Prep	PREP ;4	NOMPREP ;7
7	Sartre	NomPrXXInc	NOMPREP ;6	
8	,	Typo		
9	il	Pro	SUJ ;12	
10	le	Pro	OBJ ;12	
11	lui	Pro	PREP ;12	
12	reprocher	VCONJS		SUJ ;9, OBJ ;10, PREP ;11
13	autant que	CSub		
14	son	DetMP	DET ;15	
15	adversaire	Nom ?P		DET ;14
16	,	Typo		
17	mais	CCoord		
18	il	Pro	SUJ ;21	
19	le	Pro	OBJ ;21	
20	lui	Pro	PREP ;21	
21	reprocher	VCONJS		SUJ ;18, OBJ ;19, PREP ;20, PREP ;22
22	au nom de	Prep	PREP ;21	NOMPREP ;23
23	Sartre	NomPrXXInc	NOMPREP ;22	
24	que	CSub		COMP ;26
25	il	Pro	SUJ ;26	
26	aimer	VCONJS	COMP ;24	SUJ ;25, OBJ ;34
27	,	Typo		
28	et	CCoord		
29	qui	ProRel	SUJ ;31	
30	lui	Pro	PREP ;31	
31	ressembler	VCONJS		SUJ ;29, PREP ;30
32	,	Typo		
33	le	Det ? ?	DET ;34	
34	homme	Nom ?S	OBJ ;26	DET ;33, ADJ ;35
35	seul	Adj ? ?	ADJ ;34	
36	,	Typo		
37	le	Det ? ?	DET ;38	
38	roquentin	NomMS		DET ;37
39	.	Typo		

FIGURE 5.4 – *Exemple d'annotation par Syntex*

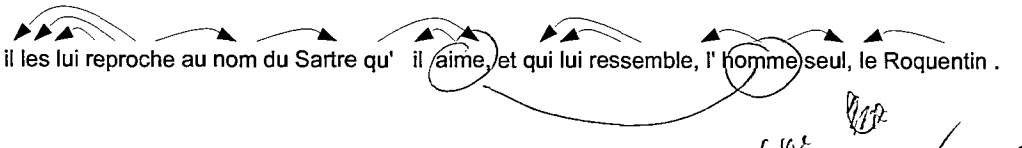


FIGURE 5.5 – *Représentation de l'annotation par Syntex*

5.3 Validation des verbes

Nos premières expériences ont révélé un nombre important d'erreurs produites par une annotation incorrecte de certains verbes par TreeTagger. Ces erreurs sont loin d'être anecdotiques car elles peuvent conduire à des « verbes » dont l'ensemble des SSC sont incorrects. Par exemple, le nom propre *Luis* est très fréquemment annoté `verbe` et lemmatisé en *luire* par TreeTagger. Mais ces erreurs ne se limitent pas aux noms propres et aux sigles. Certains mots communs posent problème : par exemple, le mot *actions* est souvent annoté `verbe` (*acter*). L'analyse syntaxique qui résulte de ces annotations est alors incorrecte. Les entrées associées à ces faux verbes peuvent être totalement fausses.

Étant donné la fréquence des erreurs observées et leur incidence sur le lexique final, il nous a semblé nécessaire de leur réserver un traitement particulier. Le repérage de ces erreurs en surface, dynamiquement lors de l'acquisition automatique a été envisagé mais il est complexe à mettre en oeuvre étant donné la variabilité des contextes provoquant ces erreurs. La solution retenue est une analyse *a priori* des verbes : nous avons cherché à répérer les mots étiquetés « verbe » alors que la majorité de leurs occurences ne correspondent pas à un verbe avant le traitement par ASSCi.

Notre méthode consiste à calculer le nombre d'occurences et la fréquence relative de chaque forme (après normalisation en minuscules) de chaque verbe dans le corpus : plus une ou deux formes prédominent pour un même mot, moins ce mot a de chances d'être un verbe.

Par exemple, toutes les occurrences du verbe *suturer* dans le corpus des 10 ans du journal *Le Monde* (voir section 6.1.1) sont issues de la forme fléchie *suture* :

```
suturer suture (1.00),
```

Dans la majorité des cas, la forme fléchie *suture* correspond au nom *suture* et a été mal analysée par TreeTagger. Par contre, pour le verbe *marcher* nous obtenons :

```
marcher marche (0.32), marchais (0.25), marchent (0.19),
marchait (0.10), marchaient (0.03), marchera (0.03),
marchons (0.02), marcherait (0.01), marchez (0.01),
marcha (0.01), marcheront (0.01), marches (0.00),
marchèrent (0.00), marchions (0.00), marcheraient (0.00),
marcherons (0.00), marcherai (0.00), marchiez (0.00),
marcherais (0.00), marchât (0.00), marcherez (0.00),
marcheras (0.00), marchai (0.00),
```

La variété des formes rencontrées confirme le statut de verbe de *marcher*.

Le programme trie ensuite les verbes dans 3 catégories, selon la fréquence relative de leur forme la plus fréquente :

- les verbes dont la fréquence de la forme la plus fréquente est supérieure à 0,99 sont exclus du processus d'acquisition ;
- les verbes dont la fréquence de la forme la plus fréquente est comprise entre 0,7 et 0,99 sont jugés par un observateur humain ;

Verbes rejetés	5049
Verbes soumis à un jugement manuel	981
Verbes gardés	5369
Total	11399
Verbes rejetés	5049
Verbes jugés incorrects par l'observateur	525
Total rejetés	5574
Verbes jugés corrects par l'observateur	456
Verbes gardés	5369
Total gardés	5825

TABLE 5.2 – *Résultats du repérage des verbes incorrects dans le corpus des 10 ans du Monde (en nombre de lemmes)*

- les verbes dont la fréquence de la forme la plus fréquente est inférieure à 0,7 sont gardés.

Ces « seuils » ont été choisis de façon empirique, suite à l'observation des verbes et de leurs formes. L'objectif est de pouvoir choisir rapidement et facilement les verbes qui seront retenus pour l'acquisition.

Dans notre exemple, la forme la plus fréquente de *marcher* est *marche* et sa fréquence relative est 0,32, il est donc classé dans la 3ème catégorie de verbes et donc gardé pour le processus d'acquisition.

Le tableau 5.2 présente le nombre de verbes classés dans chacune des catégories par le programme et les résultats du traitement après le jugement des verbes par l'annotateur humain.

Parmi les 981 verbes soumis à un jugement, 456 verbes sont jugés « corrects » par l'observateur. Ils s'ajoutent donc aux 5369 verbes gardés par le programme en premier lieu. L'acquisition se fera donc sur 5825 verbes. Au final, près de la moitié des mots annotés verbes dans le corpus sont rejetés par ce traitement. Voici quelques exemples de ces mots mal annotés (rejetés par l'algorithme ou suite à un jugement humain) :

```
senner senna (1.00),

somnambuler somnambule (1.00),

alper alpes (0.75), alpa (0.25),

caserner caserne (0.78), casernaient (0.11), casernes
(0.11),
```

Suite à ce processus, on obtient une liste des verbes sur lesquels ASSCi peut être utilisé.

Si cette approche permet d'éviter l'acquisition de SSC pour des mots qui ne sont pas des verbes, elle ne permet pas de traiter finement l'ensemble des cas d'erreurs de ce type. Il existe notamment un certain nombre de lemmes verbaux courants dont une seule forme pose problème. Par exemple, un certain nombre des occurences du verbe *marcher* sont issues du mot *Marchais*

(Georges Marchais). L'acquisition de SSC pour le verbe *marcher* est donc en partie issue de contextes erronnés. Néanmoins, le nombre total d'occurences de ce verbe dans le corpus (plus de 4000) permet de s'assurer que ces contextes n'auront pas une influence trop importante sur les entrées de *marcher* dans les SSC acquis.

5.4 Description des modules d'ASSCi

ASSCi est un système d'acquisition automatique de schémas de sous-catégorisation pour l'acquisition de SSC pour les verbes français. Cette section présente les trois modules qui le composent : l'extracteur de pré-schémas de sous-catégorisation, le constructeur de schémas candidats et le filtre de schémas non pertinents. Pour chacun de ces modules, nous décrivons son rôle dans le processus d'acquisition, nous donnons un aperçu du processus d'acquisition pour l'exemple utilisé aux sections 5.2.1 et 5.2.2, nous détaillons l'implémentation de ce module [12] et nous discutons brièvement son fonctionnement.

5.4.1 Extracteur de pré-schémas de sous-catégorisation

Le premier module extrait des « pré-schémas de sous-catégorisation » (pré-SSC) associés chacun à une occurrence d'un verbe dans le corpus à partir des sorties de l'analyseur syntaxique. Les pré-schémas comprennent les informations concernant le verbe et ses compléments (relations et étiquettes) dont le deuxième module a besoin pour former les SSC.

Par exemple, pour la phrase « *Les fameuses « erreurs » de Sartre, il les lui reproche autant que ses adversaires, mais* ***il les lui reproche au nom du Sartre qu'il aime, et qui lui ressemble, l' homme seul, le Roquentin.*** », l'extracteur produit le pré-SSC suivant :

```
      0100.anasynt!d686339p6_2!21
      REPROCHER+reprocher
      P-OBJ:SP<au nom de+SN>:Sartre | SUJ:SN:il
| OBJ:SN:le | A-OBJ:SP<à+SN>:lui
```

Le pré-schéma est composé de 3 éléments :

- `0100.anasynt!d686339p6_2!21` est l'identifiant du pré-schéma. Il est constitué du nom du fichier dont est issu le pré-SSC (`0100.anasynt`); de l'identifiant de la phrase dans le fichier (`d686339p6_2`) et de l'identifiant du verbe concerné dans la phrase (`21`). La disponibilité de cette information tout au long du processus d'acquisition permettra un retour aux données et donc une meilleure traçabilité des erreurs dans le lexique final.
- `REPROCHER+reprocher` est le verbe concerné par le pré-schéma [13].
- La suite du pré-schéma correspond aux compléments du verbe dans la phrase. Ici, *reprocher* a quatre compléments : *il*, *les*, *lui*

12. Les modules d'ASSCi sont implémentés en langage Perl.

13. Ce format facilite le classement alphabétique des verbes (en cas de caractères accentués).

Fonction		Catégorie	
SUJ	*sujet*	SN SINF	*syntagme nominal* *syntagme infinitif*
REF	*forme pronominale*	refl	*pronom*
OBJ	*objet*	SN SINF PropSub	*syntagme nominal* *syntagme infinitif* *proposition subordonnée*
ATTS	*attribut du sujet*	SA SN SINF	*syntagme adjectival* *syntagme nominal* *syntagme infinitif*
ATTO	*attribut de l'objet*	SA	*syntagme adjectival*
A-OBJ	*objet indirect régi par à*	SP<à+SN> SP<à+SINF>	*syntagme prépositionnel* *syntagme prépositionnel*
DE-OBJ	*objet indirect régi par de*	SP<de+SN> SP<de+SINF>	*syntagme prépositionnel* *syntagme prépositionnel*
P-OBJ	*objet indirect régi par la préposition* prep	SP<*prep*+SN> SP<*prep*+SINF>	*syntagme prépositionnel* *syntagme prépositionnel*

TABLE 5.3 – *Cas traités par ASSCi*

et *au nom du Sartre qu'il aime (...)*. Par exemple, le module associe `P-OBJ:SP<au nom de+SN>:Sartre` au complément *au nom du Sartre qu'il aime (...)*. Il s'agit d'un complément prépositionnel régi par *au nom de* suivi d'un syntagme nominal dont la tête nominale est *Sartre*.

Un pré-SSC est donc constitué du verbe à l'infinitif et de la liste de ses compléments. A cette étape, l'ordre des compléments n'a pas d'importance. Pour chaque complément, le module fournit le lemme de la tête, sa catégorie morphosyntaxique et sa fonction syntaxique. L'extracteur associe l'un des couples (fonction et catégorie) d'étiquettes présentés dans le tableau 5.3 à chacun des compléments du verbe.

De plus, les phrases reconnues par SYNTEX comme étant à la forme passive sont marquées par l'ajout du code `PASSIF` à la liste des compléments extraite.

Après la phase d'extraction, un pré-SSC est associé à chaque occurrence de verbe conjugué de chaque phrase du corpus. Nous avons fait le choix de ne pas acquérir de SSC à partir des verbes à l'infinitif. En effet, ces verbes apparaissent le plus souvent dans des phrases plus complexes et l'analyse effectuée par SYNTEX est alors plus incertaine. Ce choix ne pose pas de problème dans la mesure où on fournit au système une quantité de données suffisante (voir section 6.1.1).

Implémentation

Pour chaque phrase du corpus analysé par SYNTEX, l'algorithme (voir algorithme 1) repère les verbes conjugués dont le nombre d'occurrences en corpus est supérieur à un seuil fixé à 30 occurrences pour permettre les calculs statistiques lors du filtrage [14] et qui n'ont pas été rejetés lors du pré-

14. L'utilisation d'un script a au préalable permis de répertorier tous les verbes présents dans le corpus et leurs nombres d'occurrences.

traitement (voir section 5.3). Il parcourt ensuite les dépendances syntaxiques de chacun de ces verbes et cherche leur fonction syntaxique et leur catégorie. L'extracteur en déduit ensuite le pré-schéma de sous-catégorisation correspondant.

Algorithme 1: Extracteur de pré-schémas de sous-catégorisation

```
Entrées : Corpus analysé par SYNTEX
Sorties : Liste des patrons de sous-catégorisation
pour Chaque phrase faire
    pour Chaque verbe conjugué faire
        pour Chaque dépendant du verbe faire
            Cherche la fonction du complément (a) ;
            Cherche la catégorie du complément (b);
            Cherche la tête du complément (c);
        fin
        Constitue le pré-schéma de sous-catégorisation pour
        l'occurrence du verbe (d);
    fin
fin
```

Les quatre principales tâches de l'algorithme d'extraction de pré-schémas sont donc de repérer dans l'analyse syntaxique (a) la fonction syntaxique de chaque complément du verbe, (b) leur catégorie syntaxique, (c) leur lemme tête et enfin (d) de constituer un pré-SSC en exploitant ces informations.

Deux procédures permettent d'inférer la fonction syntaxique et la catégorie syntaxique de chacun des arguments. Elles utilisent des règles de généralisation qui déterminent la catégorie et la fonction syntaxique de l'argument en fonction des informations disponibles dans l'analyse de surface : relation du verbe au complément, catégorie morphosyntaxique du complément ou même les lemmes des mots qui constituent ce complément. Par exemple, si l'élément régi par le verbe par la relation `OBJ` est un nom, la catégorie du complément est `SN` et sa fonction est `OBJ`. Toutefois, dans bien des cas, l'association entre l'analyse de surface et les étiquettes de l'argument dans le pré-SSC n'est pas aussi triviale. Nous détaillons à présent le fonctionnement de ces traitements.

Une procédure est chargée de chercher les fonctions syntaxiques pour chacun des dépendants du verbe (étape (a) de l'algorithme 1). Les fonctions reconnues par l'extracteur sont données dans le tableau 5.3.

Considérons par exemple l'analyse de SYNTEX correspondant à la proposition *il les lui reproche au nom du Sartre qu'il aime* (tableau 5.4).

Le verbe *reprocher* a quatre régis :
- *il* (`18`) dont la relation au verbe est `SUJ`;
- *les* (`19`) dont la relation au verbe est `OBJ`;
- *lui* (`20`) dont la relation au verbe est `PREP`;
- *au nom du* (`22`) dont la relation au verbe est `PREP`.

Dans la plupart des cas, la fonction recherchée par l'extracteur correspond à la relation de SYNTEX : dans notre exemple, *il* est sujet de *reprocher* et *les* est son objet. Néanmoins, il existe trois exceptions notables :

18	il	Pro	SUJ ;21	
19	le	Pro	OBJ ;21	
20	lui	Pro	PREP ;21	
21	reprocher	VCONJS		SUJ ;18, OBJ ;19, PREP ;20, PREP ;22
22	au nom de	Prep	PREP ;21	NOMPREP ;23
23	Sartre	NomPrXXInc	NOMPREP ;22	
24	que	CSub		COMP ;26
25	il	Pro	SUJ ;26	
26	aimer	VCONJS	COMP ;24	SUJ ;25, OBJ ;34

TABLE 5.4 – *Analyse de* SYNTEX *correspondant à la proposition « il les lui reproche au nom du Sartre qu'il aime »*

- si le verbe est un auxiliaire, l'algorithme recommence le traitement sur le verbe conjugué. Par exemple, dans la phrase *Pierre est venu à Lille hier.*, le verbe pour lequel on cherche à acquérir un pré-SSC est le verbe *venir* (et non *être*).
- les adverbes sont exclus des pré-schémas car dans la plupart des cas, ils ne sont pas sous-catégorisables (c'est-à-dire qu'il s'agit la plupart du temps de modifieurs) et que le repérage automatique des adverbes sous-catégorisés est un problème difficile.
- la relation `PREP` donne les fonctions `A-OBJ`, `DE-OBJ` ou `P-OBJ` en fonction de la préposition tête du complément. La fonction `P-OBJ` concerne tous les compléments prépositionnels régis par une autre préposition que *à* ou *de*. Dans l'exemple utilisé dans ce chapitre, le complément *au nom du Sartre qu'il aime (...)* a pour fonction `P-OBJ` (car la relation de *au nom du* au verbe *reprocher* est `PREP` et la préposition tête est *au nom de*) alors que *lui* a pour fonction `A-OBJ` (sa relation au verbe est également `PREP` mais la préposition tête est *à*).

Une procédure est chargée de générer la catégorie syntaxique du complément ainsi que son lemme tête (étapes (b) et (c) de l'algorithme). Nous nous concentrons dans un premier temps sur la catégorie syntaxique (b). Le programme utilise essentiellement la catégorie morpho-syntaxique des mots qui composent le complément (en particulier de sa tête). Si certains cas sont simples, e.g. : *si le complément est régi par un nom, il s'agit d'un syntagme nominal*, d'autres posent des difficultés e.g. : *si le gouvernant du complément est une préposition, il faut chercher quelle est cette préposition puis identifier le type de syntagme (nominal ou infinitif) qui suit cette préposition pour obtenir une catégorie de la forme* `SP<à+SINF>`. Dans l'exemple, le complément du verbe *reprocher* d'identifiant `22` (et dont la fonction est `P-OBJ`) correspond à l'élément *au nom de* : `22 | au nom de | Prep | PREP;21 | NOMPREP;23`. Cet élément est une préposition qui régit l'élément `23` de l'analyse SYNTEX à travers la relation `NOMPREP`. Cet élément (*Sartre*) est un nom et la catégorie du complément est donc `SP<au nom de+SN>`.

Le module reconnaît les catégories suivantes :

- **Syntagmes nominaux :** si la tête du complément est un nom, un substantif ou certains pronoms.
- **Syntagmes infinitifs :** si la tête du complément est un verbe à l'infinitif.
- **Syntagmes prépositionnels :** si la tête du complément est une préposition ou certains pronoms et pronoms relatifs. Dans ce cas, le module

cherche quelle est cette préposition et quel syntagme suit cette préposition.
- **Syntagmes adjectivaux :** si la tête du complément est un adjectif ou un participe passé ou présent adjectif verbal.
- **Propositions subordonnées :** si la tête du complément est une conjonction de subordination.
- **Forme pronominale :** si la tête du complément est un pronom réfléchi ou réciproque.

Comme nous l'avons déjà vu, les compléments adverbiaux sont ignorés par le système en raison de la faible fréquence des compléments adverbiaux essentiels et de la difficulté à les repérer automatiquement à partir d'analyses de surface.

Des programmes sont chargés de traiter les cas plus complexes comme les conjonctions de coordination, les pronoms et les pronoms relatifs. Nous prendrons ici l'exemple du traitement des conjonctions de coordination qui nécessitent un traitement spécifique.

(Xa) *Bernard Kouchner, qui dirige la mission de l'ONU au Kosovo, **témoigne** de « cette très grosse erreur ou de cette manipulation ».*

Par exemple, dans la phrase (Xa), le verbe *témoigner* a pour complément *« de cette très grosse erreur ou de cette manipulation »*. Les deux compléments prépositionnels en *de* sont reliés par la conjonction de coordination *ou* qui est régie par le verbe.

La recherche des compléments reliés par une conjonction de coordination n'est pas triviale car l'annotation des cas de coordination par SYNTEX est plus ou moins précise selon la complexité de la phrase. Si la tête du complément dont on cherche à acquérir la fonction et la catégorie est une conjonction de coordination, il faut donc différencier les cas suivants :
- La catégorie du complément est « codée » dans l'étiquette de la conjonction de coordination au sein de l'analyse SYNTEX. Par exemple, dans *leur environnement économique et leur situation financière*, la conjonction de coordination *et* dont l'analyse syntaxique est `52 | et | CCoordNom | OBJ;48 | CC;50, CC;54`. La catégorie `CCoordNom` indique que *et* relie deux syntagmes nominaux. Les relations des dépendants `CC;50` et `CC;54` permet d'en déduire ces syntagmes *leur environnement économique* et *leur situation financière*. La catégorie de ce complément est donc `SN`. Les têtes nominales de ces syntagmes sont toutes les deux incluses dans le pré-SSC. La partie du pré-schéma qui correspond à ce complément est donc : `OBJ:SN:environnement,situation`.
- La catégorie de la conjonction de coordination est `CCoord` (l'analyse SYNTEX ne précise pas la catégorie des éléments coordonnés). Il faut alors extraire la catégorie et les têtes nominales des éléments coordonnés. Dans le cas où les éléments coordonnés sont de catégories différentes, il est impossible de choisir l'étiquette à associer à l'élément avec les informations dont on dispose. Aucun pré-SSC n'est alors associé au verbe en cours de traitement pour cette séquence. Il en est de même lorsque les éléments coordonnés sont régis par des prépositions

différentes (comme par exemple dans la phrase : *Mais elle diminue en Bourgogne, en Champagne Ardenne, en Auvergne et dans le Limousin.* [15]).

La « tête » (nominale ou infinitive) du complément est conservée dans le pré-schéma (c'est la procédure d'extraction de catégorie syntaxique qui est chargée d'extraire le lemme tête du complément, étape (c) de l'algorithme). Cette information n'est pas utilisée dans le processus d'acquisition (modules constructeur et filtre). Cependant, il est intéressant d'avoir ce type de données dans le lexique pour donner la possibilité aux utilisateurs du lexique d'exploiter ces informations ou pour enrichir le lexique avec des restrictions de sélection par exemple.

Après que tous les compléments du verbe ont été traités, l'extracteur constitue le pré-SSC en agglomérant ces compléments (voir algorithme 1). La liste de compléments obtenus (c'est-à-dire le pré-schéma) est non ordonnée à cette étape du processus d'acquisition.

Si la phrase est à la forme passive, cette information est incluse dans le pré-schéma de sous-catégorisation par l'extracteur. Ainsi, il sera ensuite possible d'implémenter une procédure de "dépassivation" pour traiter la phrase ou de marquer le verbe comme « passivable » dans le lexique.

Discussion

Nous venons de le montrer, les relations et catégories morpho-syntaxiques font l'objet de traitements par ce premier module. Dans une première implémentation d'ASSCi , ce traitement était effectué par le constructeur et l'extracteur récupérait les informations issues de l'analyse SYNTEX des compléments en effectuant peu de « parcours » dans l'analyse.

La deuxième implémentation des modules, qui mettait en place notamment un traitement plus élaboré des conjonctions de coordination ou des syntagmes prépositionnels, a causé la « migration » de ces traitements vers l'extracteur. L'objectif visé était d'éviter de devoir accéder à l'analyse SYNTEX après la phase d'extraction. En effet, des accès à l'analyse de l'ensemble de la phrase sont nécessaires pour traiter certaines constructions : conjonctions de coordination, syntagmes prépositionnels, etc (voir infra). Ce choix s'est avéré problématique car il implique un retour au début de la chaîne de traitements lorsqu'un changement doit intervenir dans l'algorithme (en raison de la détection d'un bug ou de l'ajout d'une nouvelle fonctionnalité par exemple). Il faut alors extraire de nouveau les pré-schémas pour l'ensemble du corpus. Or, les temps de calcul de l'extracteur sont longs car c'est le corpus tout entier qui est traité à cette étape. De plus, l'augmentation du nombre de verbes traités (par exemple en abaissant le seuil d'occurrences pour les verbes traités) augmente beaucoup le temps d'exécution de l'extracteur.

Les règles utilisées par l'extracteur peuvent être améliorées. On pourrait en effet effectuer un traitement encore plus fin des conjonctions de coordination, des pronoms ou des adverbes. Néanmoins, le parti pris de notre approche est de favoriser un traitement en largeur à un traitement en profondeur : la quantité de données doit permettre de couvrir une grande partie de la langue sans effectuer des analyses fines coûteuses et parfois impossibles

15. Extrait du corpus des 10 ans du journal Le Monde

sans l'apport de ressources extérieures. Cette approche garantit également la généricité de la méthode.

5.4.2 Constructeur de schémas de sous-catégorisation candidats

Description du module

Le constructeur de schémas utilise les informations contenues dans les pré-SSC pour « construire » les SSC candidats qui constituent un lexique de sous-catégorisation non filtré.

Contrairement à la plupart des méthodes d'acquisition de SSC employées jusqu'alors (par exemple par Preiss et al. (2007)), ASSCi ne dispose pas d'une liste de schémas prédéfinie. Ce choix s'inscrit dans notre volonté d'acquérir des schémas sans *a priori*, pour faire émerger du corpus les SSC correspondant à l'usage. Il n'y a d'ailleurs pas de raison qu'une liste de schémas existe si l'on ne dispose pas déjà d'une ressource et il est clair que la constitution d'une liste de schémas possibles pour une langue donnée n'est pas une tâche triviale. Ce choix permet même d'observer pour certains verbes des constructions qu'il était difficile de prévoir *a priori*, surtout quand on travaille sur corpus spécialisé (voir section 6.3.1).

Les SSC candidats ne concernent pas une seule occurrence en corpus mais un ensemble d'occurrences. Le constructeur a aussi pour rôle de comptabiliser le nombre d'occurrences de chaque schéma de sous-catégorisation pour chaque verbe ainsi que de calculer leur fréquence relative, c'est-à-dire la fréquence du SSC relativement au verbe. Ces informations de fréquence seront par la suite utilisées par le module de filtrage (voir section 5.4.3).

Lors des traitements du constructeur, les schémas sont normalisés. Les compléments sont ordonnés. Les compléments gouvernés par une préposition ne pouvant régir un argument sont supprimés. La liste des prépositions concernées est issue du lexique PrepLex, un lexique de prépositions du français construit en fusionnant les informations contenues dans différents lexiques disponibles[16] (Fort et Guillaume, 2007). Le module supprime également les compléments « doublons » des pré-SSC. On considère que deux compléments sont doublons lorsqu'ils sont strictement similaires par leur fonction et leur catégorie. En effet, la plupart du temps, ces doublons sont dus à des erreurs d'analyse syntaxique et les constructions contenant deux compléments similaires sont excessivement rares en français.

Le format des étiquettes utilisées pour décrire les compléments est le même que pour l'extracteur (voir section 5.4.1, tableau 5.3).

Exemple

Par exemple, la phrase « *Les fameuses « erreurs » de Sartre, il les lui reproche autant que ses adversaires, mais il les lui reproche au nom du Sartre qu'il aime, et qui lui ressemble, l' homme seul, le Roquentin.* », a produit le pré-SSC suivant :

```
      0100.anasynt!d686339p6_2!21
      REPROCHER+reprocher
      P-OBJ:SP<au nom de+SN>:Sartre | SUJ:SN:il
| OBJ:SN:le | A-OBJ:SP<à+SN>:lui
```

16. http://loriatal.loria.fr/Resources.html

Le SSC candidat issu de ce pré-schéma consolidé avec l'ensemble des pré-schémas du même type pour le verbe *reprocher* est `[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]`. Les compléments sont séparés par des virgules. Ce SSC contient trois compléments : le sujet (`SUJ:SN`), le complément d'objet (`OBJ:SN`) et l'objet prépositionnel régi par *à* (`A-OBJ:SP<à+SN>`). Remarquons que le complément prépositionnel régi par *au nom de* (`P-OBJ:SP<au nom de+SN>`) n'apparaît pas dans le schéma candidat (le constructeur a supprimé ce complément dont la préposition (*au nom de*) est non argumentale).

Le constructeur a trouvé 2128 occurences de cette entrée dans le corpus, ce qui produit une fréquence relative de 0.218 (le nombre total d'occurences de *reprocher* dans le corpus étant de 9757). Le module attribue également un identifiant à l'entrée (ici `2610`) et conserve les identifiants des phrases correspondant à cette entrée dans le corpus ainsi que les lemmes têtes des différents arguments.

Implémentation

L'algorithme 2 présente les opérations effectuées par le constructeur pour passer d'une liste de pré-SSC à une liste de SSC globaux.

Algorithme 2: Constructeur de schémas candidats

Entrées : Liste des pré-schémas de sous-catégorisation
Sorties : Liste de SSC non filtrée
pour *Chaque pré-schéma de sous-catégorisation* **faire**
  Ordonne les compléments ;
  Exclut les compléments régis par des prépositions non sous-catégorisables et les compléments doublons ;
  Crée un nouveau SSC ou ajoute une occurrence au SSC s'il existe ;
fin
pour *Chaque schéma de sous-catégorisation* **faire**
  Calcule la fréquence relative du SSC ;
fin

Afin de pouvoir généraliser les pré-SSC en SSC, il faut ordonner la liste des compléments. Considérons par exemple les phrases suivantes :

(a) *Lucas donne un livre à Martine.*
(b) *Lucas lui donne un livre.*
(c) *A qui Lucas a-t-il donné un livre ?*

Ces phrases diffèrent par leur structure de surface mais rendent compte du même comportement syntaxique profond. Il est donc souhaitable qu'elles produisent le même SSC. Les schémas produits par ces trois phrases à partir des pré-schémas diffèrent en raison de l'ordre des compléments dans la phrase : `[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]` pour (a), `[SUJ:SN, A-OBJ:SP<à+SN>, OBJ:SN]` pour (b) et `[A-OBJ:SP<à+SN>, SUJ:SN, OBJ:SN]` pour (c). En ordonnant ces schémas, on obtient un SSC unique : `[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]` en traitant ces trois phrases.

Les éléments des SSC sont donc ordonnés selon leur fonction, dans l'ordre suivant : sujet, forme pronominale (étiquette REF), objet, groupe prépositionnel gouverné par *à*, groupe prépositionnel gouverné par *de*, autres groupes prépositionnels, attribut du sujet, attribut de l'objet[17]. Cette normalisation permet de s'affranchir des réalisations de surface. En effet, les phrases contenant des pronoms, des inversions ou des extrapositions n'ont pas une structure canonique et l'on souhaite s'abstraire de ces variations de surface.

Les compléments gouvernés par une préposition ne pouvant régir un argument sont ensuite supprimés. La liste de ces prépositions est issue du lexique PrepLex, un lexique de prépositions du français[18] Fort et Guillaume (2007). Dans PrepLex, les prépositions sont classées en différentes catégories. L'une de ces catégories permet de distinguer les prépositions argumentales des prépositions non-argumentales. La liste des prépositions argumentales issue de PrepLex est donnée en annexe C.

Pour chaque complément prépositionnel, le constructeur cherche dans la liste de prépositions argumentales si la préposition qui le régit est susceptible de porter la sous-catégorisation ou non. Dans le cas où la préposition est non argumentale, le complément est retiré du schéma. Dans l'exemple utilisé dans ce chapitre, le complément régi par la préposition *au nom de* (`P-OBJ:SP<au nom de+SN>:Sartre`) est exclu du SSC obtenu en sortie du constructeur (`[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]`).

Cette technique permet d'éliminer certains compléments prépositionnels modifieurs évidents des SSC mais est bien sûr insuffisant pour règler le problème de la distinction argument-modifieur qui dépend essentiellement du contexte et non de la préposition qui régit le complément.

Le module supprime également les compléments « doublons » des pré-SSC. On considère que deux compléments sont doublons lorsqu'ils sont strictement similaires par leur fonction et leur catégorie.

Observons par exemple, la phrase :

(a) *Douze millions de téléspectateurs ont suivi, en France, dimanche 12 décembre, la finale du Championnat du monde de handball féminin qui a opposé l'équipe de France à la Norvège à Lillehammer.*

Le verbe *opposer* a ici deux compléments prépositionnels régis par *à* : *à la Norvège* et *à Lillehammer*. Le SSC construit correspondant à cette phrase est `[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>, A-OBJ:SP<à+SN>]` qui contient un doublon de compléments prépositionnels régis par *à*. L'un de ces compléments est retiré du SSC, ce qui permet de supprimer un modifieur (*à Lillehammer*) du schéma. Le schéma résultant du traitement est donc `[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]`. Toutefois, un SSC qui contient deux compléments régis par la même préposition mais dont la catégorie du lemme tête est différente (comme par exemple `[SUJ:SN, A-OBJ:SP<à+SN>, A-OBJ:SP<à+SINF>]`) ne sera

17. Ce classement a également son importance pour l'implémentation de la réduction des SSC à l'étape de filtrage (voir section 5.4.3)

18. `http://loriatal.loria.fr/Resources.html`

pas modifié puisque les deux compléments en *à* (A-OBJ:SP<à+SN> et A-OBJ:SP<à+SINF>) ne sont pas strictement similaires.

Après la normalisation et la constitution des SSC, le constructeur de SSC calcule le nombre d'occurrences de chaque couple verbe-schéma et sa fréquence relative.

La fréquence relative d'un couple est calculée en faisant le rapport entre le nombre d'occurrence du couple verbe-schéma en corpus et le nombre d'occurrences du verbe :

$$freq\_rel(verbe_i, ssc_j) = \frac{|verbe_i, ssc_j|}{|verbe_i|}$$

Par exemple, la fréquence relative de l'entrée correspondant au verbe *reprocher* et au schéma [SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>] est 0.218. Ce chiffre signifie que près de 22 % des constructions du verbe *reprocher* dans notre corpus produisent le schéma [SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]. La somme des fréquences relatives des entrées correspondant à un même verbe est donc égale à 1.

Discussion

La normalisation (ordonnancement des compléments, suppression des compléments prépositionnels non argumentaux ou des compléments doublons) effectuée lors de l'étape de construction des schémas est discutable. En effet, le dosage entre normalisation et perte d'informations est délicat. Il est cependant aisé de faire des modifications dans le programme pour désactiver l'un ou l'autre de ces traitements.

De plus, on sait que les compléments modifieurs ont plus tendance à être déplaçables et on pourrait utiliser l'ordre des compléments pour différencier arguments et modifieurs. Néanmoins, une normalisation de l'ordre des SSC permet d'éviter l'éclatement de constructions très proches (un tel éclatement serait nocif pour le filtrage). Un bon compromis serait de garder tous les ordres observés en corpus dans un champ séparé du SSC, ce qui permettrait de normaliser le SSC sans perte d'informations.

5.4.3 Filtre des schémas non pertinents

Description du module

Tous les systèmes d'acquisition nécessitent une phase de filtrage à la fin du processus d'acquisition. En effet, les schémas candidats extraits par le constructeur sont bruités car ils sont parfois construits à partir d'étiquetages ou d'analyses syntaxiques incorrects. Il se peut également qu'une règle de construction produise des SSC incorrects à cause d'une généralisation ou d'une normalisation trop importante. Enfin, la présence de modifieurs dans les schémas génère également des erreurs que l'on souhaite filtrer (Manning, 2003) (voir chapitre 2.3.2).

Le filtre examine donc les SSC candidats issus du constructeur et compare leurs fréquences relatives à un ou plusieurs seuils déterminés par un étalonnage automatique du système (technique de l'estimation du maximum de vraissemblance, voir *infra*). Si la fréquence du schéma est supérieure au

seuil, il « passe » le filtre et figurera dans le lexique. Dans le cas où la fréquence du SSC est inférieure au seuil, il est rejeté par le module à moins qu'il soit possible de réduire ce schéma en supprimant un de ses compléments prépositionnels et en passant le SSC résultant au filtre (ce traitement est détaillé dans l'implémentation). La sortie de ce module est une liste d'entrées (paires verbe - SSC) qui constituent le lexique de sous-catégorisation.

Exemple

Par exemple, le SSC `[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]` résultant notamment de la phrase « *Les fameuses « erreurs » de Sartre, il les lui reproche autant que ses adversaires, mais il les lui reproche au nom du Sartre qu'il aime, et qui lui ressemble, l' homme seul, le Roquentin.* » est soumis au filtrage.

Sa fréquence relative (0.22) est comparée au seuil :
- si cette fréquence est supérieure au seuil, le schéma apparaîtra dans le lexique pour le verbe *reprocher* ;
- si cette fréquence est inférieure au seuil, le complément prépositionnel (`A-OBJ:SP<à+SN>`) est supprimé et le SSC `[SUJ:SN, OBJ:SN]` est de nouveau testé par le filtre après un nouveau calcul de son nombre d'occurences et de sa fréquence relative.

ASSCi utilise un seuil de 0.06 et l'entrée `reprocher - [SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]` passe l'étape de filtrage.

Implémentation

L'algorithme 3 présente les traitements effectués par le module de filtrage.

Algorithme 3: Filtre des schémas non pertinents

Entrées : Liste de SSC candidats (non filtrée)
Sorties : Liste de SSC (filtrée)
pour *Chaque entrée* **faire**
  si *Fréquence relative(entrée)* >= *seuil* **alors**
    Ajoute l'entrée au lexique ;
  sinon
    Retire un des compléments prépositionnels du schéma;
    Consolide l'entrée qui en résulte;
    Cette entrée est de nouveau testée par le filtre;
  fin
fin

La technique de filtrage retenue pour ASSCi est un seuil sur les fréquences relatives des entrées qui sont calculées par le constructeur (voir section 5.4.2). Afin de déterminer le meilleur seuil, on étalonne le système en utilisant le lexique TreeLex comme référence (le détail de l'étalonnage d'ASSCi est donné plus bas).

Cette méthode, appelée maximum de vraisemblance (*maximum likelihood estimates*), est décrite dans (Korhonen et al., 2000) qui la compare à deux

autres techniques de filtrage : la loi binomiale, souvent utilisée (*binomial hypothesis testing*) (Brent, 1993; Manning, 1993) et la log-vraisemblance (*binomial log-likelihood ratio* (Gorrell, 1999)). Comme nous l'avons vu dans la section 4.4, le maximum de vraisemblance donne les meilleurs résultats. C'est pourquoi nous avons décidé d'utiliser cette méthode.

Pour chacune des entrées du lexique non filtré, sa fréquence relative est comparée à un seuil déterminé par les expériences d'étalonnage décrites plus bas. Si la fréquence relative est supérieure au seuil, l'entrée est retenue dans le lexique de sortie ; sinon, l'entrée est rejetée. La valeur du seuil utilisé dans le filtre est de 0,06.

Afin d'améliorer cette méthode, nous avons déterminé des seuils différenciés pour les SSC ne contenant qu'un sujet et les schémas avec un marqueur de forme pronominale. En effet, lorsque le rattachement d'un élément de la phrase à un autre est incertain, l'implémentation de SYNTEX privilégie de ne pas lier les éléments. Certaines phrases complexes (incises, ...) produisent des analyses erronées et le schéma résultant de cette analyse est souvent `SUJ:SN`. Le seuil utilisé pour ce schéma est donc fixé à 0,09 par le processus d'étalonnage. Par ailleurs, les différents types de compléments pronominaux sont difficiles à distinguer (comme nous l'avons montré dans le chapitre 2). Pour réduire les erreurs liées à cette difficulté, le seuil pour ces schémas pronominaux est de 0,22.

Nous avons également mis des seuils sur le nombre d'occurences des verbes et des SSC. En effet, le filtrage sur la fréquence relative ne permet pas d'éviter qu'un verbe très peu présent en corpus produise des entrées erronées. Par exemple, le verbe *ondoyer* a 10 occurrences dans le corpus et il produit (après filtrage) 7 SSC dont la plupart n'apparaissent qu'une ou deux fois. Le schéma `[SUJ:SN,P-OBJ:SP<avec+SN>]` apparaît une seule fois. Or, la phrase qui a produit cette entrée est *La Danse d' Anitra ondoie avec sensualité.* Sa fréquence relative (0,100) est supérieure au seuil en raison du faible nombre d'occurrences de *ondoyer*. Cependant, le complément `P-OBJ:SP<avec+SN>` est manifestement un modifieur et l'entrée devrait être filtrée. La mise en place d'un seuil de 30 sur le nombre d'occurrences du verbe et de 10 sur le nombre d'occurrences du schéma permet d'éviter ce problème.

Un grand nombre d'erreurs est lié à la problématique de la distinction argument/modifieur. Comme dans la plupart des travaux existants (Briscoe et Carroll, 1997; Gardent, 2009), nous comptons sur le fait que les SSC comprenant des modifieurs sont moins fréquents en corpus que les SSC comprenant uniquement des arguments. L'un des rôles de l'étape de filtrage est donc de rejeter les schémas incluant des modifieurs. Toutefois, l'information perdue en rejetant ces schémas peut être utile et produire d'autres SSC.

Par exemple, la phrase « *Ce RDS s'applique d'abord comme la CSG aux revenus d'activité des salariés et des non-salariés.* » produit le SSC `[SUJ:SN, REF:refl, A-OBJ:SP<à+SN>, P-OBJ:SP<comme+SN>]` pour le verbe *appliquer*. Ce schéma, trop peu fréquent dans le corpus, est rejeté par le filtre. Pourtant, en supprimant le modifieur `P-OBJ:SP<comme+SN>`, on obtient un SSC valide : `[SUJ:SN, REF:refl, A-OBJ:SP<à+SN>]`.

Nous avons donc amélioré le module en ajoutant un processus de « réduction » des schémas contenant un modifieur supposé. Plutôt que d'exclure totalement ces SSC du lexique, nous avons mis en place un processus de réduction des schémas : lorsqu'un schéma contenant au moins un complément prépositionnel est rejeté par le filtrage, on essaie de le ramener à un schéma moins complexe, c'est-à-dire sous-catégorisant un complément prépositionnel de moins. On espère ainsi éliminer un modifieur et ramener le schéma à un SCC valide pour le verbe (sur la base de sa fréquence relative). L'entrée résultante est ensuite « consolidée » avec le nombre d'occurrences du schéma contenant le modifieur, ce qui permet d'augmenter la probabilité pour ce SSC de passer l'étape de filtrage. Les schémas sont traités par ordre décroissant de leur longueur (i.e. leur nombre de compléments), ce qui permet d'assurer que les schémas dont la fréquence relative est augmentée par des schémas réduits sont traités après ceux susceptibles d'augmenter cette fréquence.

Par exemple, la phrase *Jean boit un café à la terrasse* permet d'identifier un SSC où le groupe prépositionnel introduit par *à* est un complément potentiel, mais comme la fréquence de ce SSC est inférieure au seuil, on peut réduire le schéma pour identifier ici un emploi transitif `[SUJ:SN, OBJ:SN]` du verbe *boire*.

Enfin, ce module recalcule également les fréquences relatives des entrées en fonction du nombre d'occurrences des verbes résultant du filtrage. En effet, le rejet de schémas réduit le nombre d'occurrences des verbes et la somme des fréquences relatives d'un verbe doit être égale à 1. Pour l'entrée présentée ci-dessus (schéma `[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]` pour le verbe reprocher), la fréquence relative finale est de 0,256.

Étalonnage du filtre

La qualité du filtrage (et *in fine* de l'acquisition) dépend des seuils utilisés. Nous avons donc défini un protocole pour choisir les seuils les plus efficaces pour le filtrage des SSC candidats. Nous avons développé un script qui permet de calculer la proximité des entrées acquises par notre système avec TreeLex en fonction du seuil avec lequel il est filtré. Si TreeLex ne peut pas être utilisé comme un *gold standard* pour l'évaluation (voir chapitre 6.2), il s'agit d'une ressource acquise à partir d'un corpus annoté « à la main », ce qui en fait un bon modèle pour l'étalonnage.

On calcule la proximité d'un lexique avec TreeLex en utilisant la F-mesure. Toutefois, nous n'utilisons pas la F-mesure habituelle (F_1) qui donne le même poids à la précision et au rappel. TreeLex a été acquis à partir d'un corpus dont la taille ne permet pas d'assurer une couverture de l'ensemble de la langue. C'est pourquoi nous privilégions le rappel sur la précision en fixant β à 1,5 (le rappel pèse alors 1,5 fois plus que la précision sur le score obtenu). Cela permet de compenser le fait qu'un déficit de précision résulte ici assez souvent d'une insuffisance de TreeLex (nous montrons dans la section 6.2 qu'une part importante d'entrées nouvelles acquises par ASSCi par rapport à TreeLex est correcte).

$$F_\beta = (1+\beta^2) * \frac{Precision * Rappel}{\beta^2 * Precision + Rappel}$$

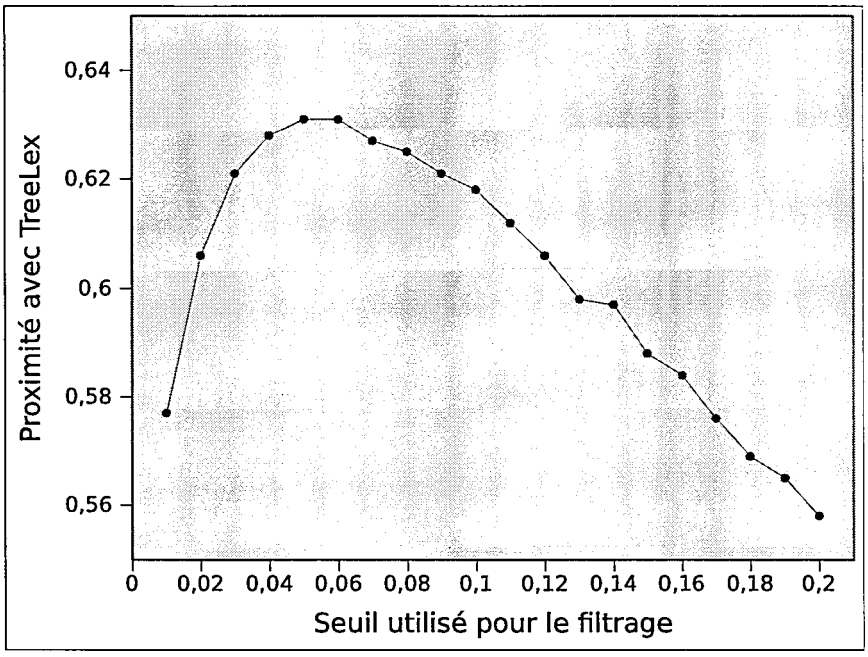


FIGURE 5.6 – *Variation de la proximité avec* TREELEX *($F_{1,5}$) en fonction du seuil*

La mesure utilisée est donc :

$$F_{1,5} = 3,25 * \frac{Precision*Rappel}{2,25*Precision+Rappel}$$

Une première expérience a permis de déterminer le meilleur seuil pour notre système (sans seuils différenciés en fonction des SSC). Nous avons donc utilisé notre script pour étudier les variations de la proximité de la ressource filtrée avec TREELEX en fonction du seuil utilisé. La figure 5.6 montre le résultat obtenu.

La plus haute $F_{1,5}$ (F-mesure avec β à 1,5) est obtenue pour un seuil de 0,06.

Une deuxième expérience a permis de déterminer les seuils les plus performants (en utilisant la même méthode que précédemment) lorsque l'on prend en compte les particularités des SSC intransitifs ou pronominaux. En faisant varier trois seuils (seuil « par défaut », seuil pour les constructions intransitives et seuil pour les constructions pronominales), nous avons étalonné le système pour qu'il utilise la meilleure combinaison possible de ces seuils. Le tableau 5.5 donne un aperçu des résultats obtenus pour quelques combinaisons de seuils.

À l'issue de l'étalonnage, les seuils utilisés dans ASSCi sont donc :
- 0,06 par défaut ;
- 0,09 pour les constructions intransitives ;
- 0,22 pour les constructions pronominales.

Discussion

L'étape de filtrage est cruciale pour les systèmes d'acquisition automatique de SSC : il faut trouver le meilleur équilibre possible entre le rejet

Seuil par défaut	Seuil intransitives	Seuil pronominales	Précision	Rappel	F-1,5
0,02	0,10	0,10	0,451	0,733	0,615
0,04	0,14	0,22	0,559	0,646	0,616
0,06	**0,09**	**0,22**	**0,584**	**0,651**	**0,629**
0,06	0,10	0,22	0,591	0,646	0,628
0,07	0,10	0,22	0,601	0,635	0,624
0,08	0,10	0,19	0,605	0,632	0,623
0,09	0,10	0,19	0,612	0,625	0,621
0,10	0,11	0,25	0,633	0,605	0,613
0,12	0,18	0,30	0,686	0,555	0,590

TABLE 5.5 – *Résultats obtenus pour quelques combinaisons de seuils lors de l'étalonnage*

d'informations erronées (afin d'améliorer la précision) et le fait de garder un maximum d'informations correctes (afin de maintenir un bon taux de rappel). En effet, il se peut que des entrées peu fréquentes dans le corpus (et donc rejetées par le filtre) soient correctes tout comme il est possible que des entrées fréquentes en corpus soient incorrectes (par exemple en raison de la présence de modifieurs). Les gains de performances acquis en optimisant le réglage des techniques de filtrage sont relativement faibles. Au delà d'un certain seuil, seul l'ajout d'informations extérieures (croisement avec des lexiques existants, « lissage » des SSC, etc) permet une amélioration significative des résultats (Korhonen et al., 2006).

D'autres techniques de filtrage (test binomial, ...) pourraient être explorées afin de vérifier si l'équivalence relative des méthodes observée dans (Korhonen et al., 2000) se vérifie pour le français. Toutefois, on a pu voir au chapitre 4 que le test binomial avait déjà été utilisé pour l'acquisition de SSC pour le français (Chesley et Salmon-Alt, 2006) sans que les résultats obtenus en F-Mesure soient très différents de ceux d'ASSCi (Messiant, 2008).

La réduction des schémas risque de produire des schémas faux. L'expérience présentée au chapitre 6 montrera la différence de résultats entre l'acquisition d'un lexique en utilisant la réduction de schémas et le filtrage sans réduction pour vérifier l'apport de ce traitement.

5.5 CONCLUSION

Nous avons développé ASSCi, un système d'acquisition automatique de SSC pour les verbes du français. Si ce système repose sur le schéma global des techniques d'acquisition présenté au chapitre 4, il explore des pistes nouvelles ou peu explorées pour l'acquisition de SSC. Tout d'abord, un prétraitement simple permet de réduire le bruit en repérant les verbes mal annotés dans le corpus (section 5.3). De plus, les SSC sont inférés directement du corpus et le système ne dispose pas d'une liste de schémas, ce qui facilite l'émergence de comportements qui ne sont pas répertoriés dans les ressources existantes (section 5.4.2). Enfin, nous avons introduit la réduction des SSC rejetés par le filtre par suppression d'un complément prépositionnel afin de mieux prendre en compte la distinction argument - modifieur (section 5.4.3).

ASSCi a permis l'acquisition d'un lexique de SSC à large couverture pour

le français appelé LexSchem. La section suivante présente ce lexique, son format et les outils qui y sont associés.

6

Utilisation et évaluation des méthodes d'acquisition automatique de SSC

Les méthodes d'acquisition automatique de SSC à partir de corpus ont été beaucoup étudiées depuis les années 90 (voir chapitre 4). Si l'intérêt de ces méthodes est indiscutable lorsqu'elles sont utilisées pour des langues ne disposant pas de lexiques de sous-catégorisation, il semble au premier abord moins pertinent lorsque la langue étudiée dispose déjà d'un certain nombre de ressources (c'est notamment le cas du français comme en témoigne l'état de l'art dressé au chapitre 3). En effet, la qualité des ressources acquises automatiquement n'atteint pas encore celle des dictionnaires développés depuis de nombreuses années dans les laboratoires et équipes de linguistique.

Toutefois, ces techniques d'acquisition sont complémentaires des ressources développées « à la main ». Elles permettent notamment d'acquérir des probabilités d'emploi des schémas et sont adaptables (elles permettent d'acquérir facilement des informations lexicales sur des sous-domaines de la langue).

Ce chapitre présente des expériences et des outils qui montrent l'apport de ces ressources : dans un premier temps, nous avons acquis un lexique de sous-catégorisation à partir d'un corpus journalistique ; nous avons ensuite évalué ce lexique ; d'autres expériences ont permis de mettre en évidence l'adaptabilité des méthodes d'acquisition automatiques ; des outils ont enfin été développés pour paramétrer la ressource obtenue et la rendre plus efficace en fonction de la tâche visée.

6.1 Acquisition d'un lexique de sous-catégorisation à large couverture pour le français

Nous avons réalisé une expérience d'acquisition automatique d'un lexique de sous-catégorisation pour le français. La ressource obtenue par cette expérience s'appelle LexSchem. Cette section présente tout d'abord le corpus utilisé pour acquérir notre lexique. Nous détaillons ensuite les réglages utilisés dans ASSCi pour l'acquisition de LexSchem. La section se conclut par une présentation du lexique et de l'interface qui permet de le consulter en ligne.

6.1.1 Matériel et méthode

Corpus utilisé

Les informations acquises lors d'une expérience d'acquisition automatique de SSC dépendent directement du corpus utilisé. Le choix de ce corpus est donc crucial et détermine la qualité et l'exhaustivité de la ressource acquise. Le choix du corpus dépend aussi de l'utilisation qui sera faite de la ressource (domaine visé, nombre de verbes concernés, etc).

Nous avons choisi d'utiliser le corpus journalistique LM10. Ce corpus est constitué des articles de 10 années (1991-2000) du journal *Le Monde*. Il s'agissait du plus gros corpus disponible et homogène quant à son style pour le français au début de nos travaux : il contient 9,8 millions de séquences (chaque séquence peut contenir une ou plusieurs phrases [1]) soit 200 millions de mots.

L'utilisation d'un corpus journalistique garantit une certaine hétérogénéité des thèmes abordés et des constructions syntaxiques utilisées du fait de la variété des auteurs, les citations présentes dans les articles, etc. Les trois phrases suivantes, extraites du corpus, montrent bien la variété des thèmes abordés et des constructions observables dans *Le Monde* ainsi que la complexité des phrases du corpus :

Le marché de New-York ne cède plus à la panique : le 2 janvier en fin de séance, les cours du brut avaient sérieusement rechuté, perdant 2 dollars par baril.

Le sida leur « bouffe la vie » : « Impossible de ne pas y penser ».

De son côté, le porte-parole du RPR, Patrick Devedjian, a déclaré que « les communiqués d'autosatisfaction, les communiqués sur les lendemains qui chantent, les Français commencent à s'en fatiguer, car les impôts, il y en a trop ».

Le corpus utilisé a été normalisé et « nettoyé » grâce à des scripts développés par Benoît Habert (afin d'éviter par exemple que les titres soient « collés » au corps des articles, etc). Ces traitements permettent, par exemple, de limiter le nombre d'erreurs dues à une mauvaise segmentation.

La distribution des verbes dans le corpus est représentée sur une échelle double-logarithmique [2] dans la figure 6.1. Cette distribution est de type zipfien (c'est-à-dire qu'il y a beaucoup de verbes très peu fréquents (quelques

1. Ce phénomène est dû à des erreurs de segmentation.

2. L'échelle logarithmique permet de mieux représenter une gamme plus étendue de valeurs (ici par exemple, le nombre d'occurrences des verbes varie de 1 à plus de 3,5 millions).

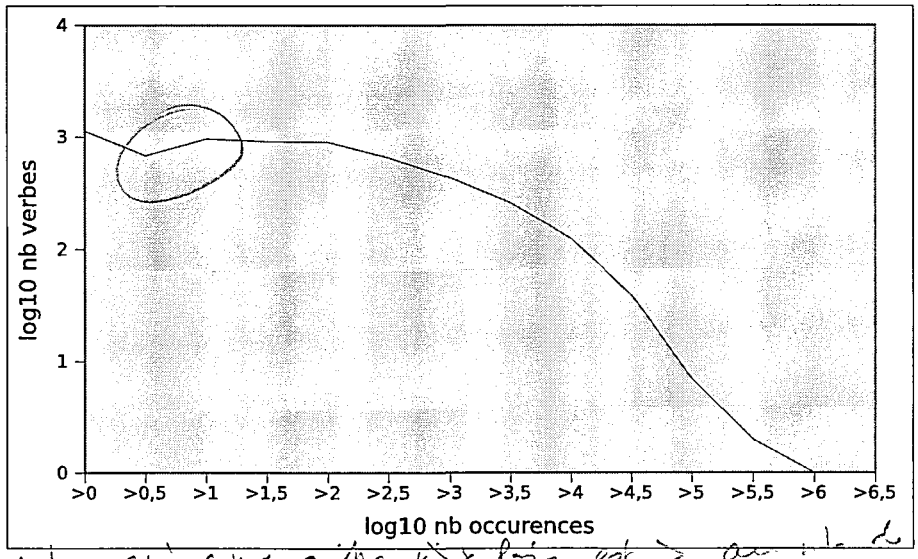


FIGURE 6.1 – *Représentation log-log de la distribution des verbes du corpus LM10*

occurrences) et très peu de verbes très fréquents (plus d'un million d'occurrences) dans le corpus). Le tableau 6.1 donne le nombre d'occurrences dans le corpus LM10 de 20 verbes.

Réglages

Nous avons utilisé le système ASSCi sur le corpus LM10 pour acquérir un lexique de sous-catégorisation verbale.

L'acquisition de LEXSCHEM a été réalisée à l'aide de la version d'ASSCi présentée dans le chapitre 5[3].

Les seuils utilisés pour le filtrage du lexique sont donc ceux déterminés par l'étalonnage présenté à la section 5.4.3 :

- 0,10 pour les schémas intransitifs ;
- 0,22 pour les formes pronominales ;
- 0,06 pour tous les autres SSC.

Rappelons également que seuls les 4 632 verbes gardés lors de la validation des verbes présentée à la section 5.3 ont été traités par ASSCi. Les verbes mal annotés par TREETAGGER sont donc exclus du processus d'acquisition.

Avant l'étape de filtrage, ASSCi répertorie 173 870 entrées différentes (couples verbe - schéma candidat). Ce chiffre témoigne de la quantité et de la variété des informations présentes dans le corpus. Si une grande partie de ces entrées sont filtrées, il est intéressant de disposer de données classées par usage syntaxique : des traitements simples sur ces données permettent de consulter par exemple les phrases du *Monde* où tel verbe apparaît avec telle préposition. Outre le traitement en largeur présenté dans cette thèse,

3. Des versions intermédiaires de LEXSCHEM ont été rendues publiques tout au long du travail de thèse. La version présentée dans ce chapitre est la version disponible sur la page web du lexique (en juin 2010) : `http://www-lipn.univ-paris13.fr/~messiant/lexschem.html`.

avoir	3503441
être	2808658
pouvoir	446981
devoir	352458
faire	236004
venir	113175
permettre	75599
prendre	53724
écrire	32457
accompagner	9601
aider	6726
acquérir	1978
remercier	888
catastropher	274
somnoler	128
nationaliser	70
importuner	29
poireauter	14
braire	3
lobotomiser	1

TABLE 6.1 – *20 verbes et leurs fréquences dans le corpus LM10 (verbes identifiés à l'aide de l'étiquette donnée par* TREETAGGER*)*

on peut donc envisager un usage des informations acquises par ASSCi dans des analyses en profondeur.

Temps de traitement pour le corpus LM10

Le temps de calcul observé sur une machine récente pour le traitement du corpus *LM10* est d'environ 30 heures pour l'extracteur de pré-schémas. Il est plus raisonnable pour les deux autres modules : un peu plus d'une heure pour le constructeur de schémas et moins de 2 minutes pour le filtrage.

Le temps de calcul élevé de l'extracteur est principalement dû à la taille des fichiers traités (9,3 gigaoctets pour le corpus LM10 analysé par SYNTEX). Il peut être considérablement réduit en ne conservant pas certaines informations volumineuses comme la liste des identifiants des analyses phrases du corpus ou les lemmes têtes des arguments.

6.1.2 Présentation du lexique

L'utilisation d'ASSCi avec les réglages décrits dans la section précédente a permis d'acquérir le lexique LEXSCHEM.

Après le filtrage des 173 870 couples verbe - schéma candidat issus du constructeur, LEXSCHEM contient 7 239 couples verbe-SSC répartis sur 3 123 lemmes verbaux et 88 SSC différents (la liste des SSC présents dans LEXSCHEM est donnée en annexe E). Environ 96 % des entrées sont donc filtrées par le troisième module. Cette proportion très importante d'entrées filtrées peut surprendre au premier abord. Elle s'explique pourtant assez simplement :

Nb de SSC	Nb de verbes	Proportion	Exemples de verbes
1	626	20,0%	*éternuer, miroiter, jubiler*
2	1385	44,3%	*accompagner, trier, retrouver*
3	729	23,3%	*ouvrir, acheter, plaire*
4	278	8,9%	*mordre, négocier, obliger*
5	88	2,8%	*inviter, jeter, manquer*
6	15	0,5%	*varier, conseiller, demander*
7	2	0,1%	*travailler, débouler*

TABLE 6.2 – *Nombre de verbes par nombre de SSC*

travailler	débouler
[SUJ:SN]	[SUJ:SN]
[SUJ:SN,OBJ:SN]	[SUJ:SN,OBJ:SN]
[SUJ:SN,A-OBJ:SP<à+SN>]	[SUJ:SN,A-OBJ:SP<à+SN>]
[SUJ:SN,P-OBJ:SP<avec+SN>]	[SUJ:SN,P-OBJ:SP<en+SN>]
[SUJ:SN,P-OBJ:SP<sur+SN>]	[SUJ:SN,P-OBJ:SP<sur+SN>]
[SUJ:SN,P-OBJ:SP<dans+SN>]	[SUJ:SN,P-OBJ:SP<dans+SN>]
[SUJ:SN,P-OBJ:SP<pour+SN>]	[SUJ:SN,DE-OBJ:SP<de+SN>]

TABLE 6.3 – *Entrées de* LEXSCHEM *pour les verbes travailler et débouler (après filtrage)*

- le nombre de verbes traités passe de 4 632 à 3 123. Un tiers des verbes (et donc toutes les entrées qui leur sont associées) sont donc supprimés du lexique final en raison d'un trop faible nombre d'occurrences (moins de 30) ;
- le seuil sur le nombre d'occurrences des SSC (toutes les entrées avec moins de 10 occurrences sont filtrées) réduit également de beaucoup le nombre d'entrées avant le filtrage par fréquence relative ;
- un nombre très important d'entrées sont filtrées parce que leur fréquence relative dans le corpus ne dépasse pas les seuils. La majorité de ces entrées sont des constructions qui contiennent un modifieur ou qui résultent d'erreurs d'analyse. Nous verrons plus loin que certaines d'entre elles sont des usages corrects mais qui n'apparaissent pas suffisamment dans le corpus *LM10* pour passer le filtrage.

Le nombre moyen de SSC par verbe est d'environ 2,3. Le nombre de SSC par verbe varie entre 1 et 7 (le tableau 6.2 donne la répartition des verbes du corpus en fonction de leur nombre de schémas).

On remarque que les verbes qui ont 1, 2 ou 3 schémas représentent plus de 90% des verbes du lexique. Seuls 0,6% des verbes acceptent 6 ou 7 SSC.

Les verbes acceptant 7 SSC sont : *travailler* et *débouler*. Le tableau 6.3 donne la liste des entrées de ces verbes.

Les entrées du verbe *travailler* correspondent bien à des usages de ce verbe : *travailler sur sa thèse, travailler dans l'informatique, ...* Toutefois, le complément `P-OBJ:SP<avec+SN>` serait sûrement considéré comme étant un modifieur par un annotateur humain dans le SSC `[SUJ:SN,P-OBJ:SP<avec+SN>]` (*travailler avec ses collègues*). Cette construction ne figure d'ailleurs pas dans DICOVALENCE[4]. On remarquera

4. Notons toutefois que ce type d'information peut permettre de repérer ultérieurement des classes sémantiques en se fondant sur les familles de verbes partageant les mêmes modi-

enfin que notre méthode ne repère pour le verbe *travailler* aucun cas de composition de compléments, bien que plusieurs compléments prépositionnels seraient susceptibles d'apparaître ensemble (*il travaille avec ses collègues sur un projet*).

Par contre, les entrées du verbes *débouler* contiennent de nombreuses erreurs. La plupart de ces erreurs résultent du faible nombre d'occurrences de ce verbe dans le corpus (environ 300) et de constructions non standard. Ainsi, le schéma `[SUJ:SN,OBJ:SN]` est le plus souvent dû à des erreurs d'analyse de SYNTEX en raison de très nombreuses inversions du sujet comme dans la phrase « *Chaque matin déboule la petite charrette de socca* » (où petite charrette est analysé comme un complément d'objet et non comme le sujet). Le SSC `[SUJ:SN,P-OBJ:SP<en+SN>]` résulte d'expressions figées ou semi-figées comme « *débouler en masse* » ou « *débouler en trombe* ». Il s'agit d'erreurs dans la mesure ou « *en trombe* » et « *en masse* » font partie d'expressions figées et ne constituent pas des compléments en tant que tels : il serait sans doute possible d'aller plus loin en affinant le processus d'acquisition. Un calcul de cooccurrence plus fin (concernant la force du lien entre le nom complément et le verbe) permettrait sans doute de repérer au moins partiellement les expressions figées et de les écarter du processus d'acquisition (en s'inspirant des travaux de (Fabre et Bourigault, 2008) – sauf qu'il s'agit ici de modéliser le continuum entre compléments « libres » et expression figées – ou bien encore de ceux de (de Cruys et Moirón, 2007)). Faute de temps, ce processus n'a pu être mis en place dans la thèse mais constitue une perspective à ce travail.

On voit ici tout ce que le processus de filtrage a d'imparfait : il faudrait éliminer les SSC qui ont des compléments fréquents qui sont en fait des modifieurs ou qui forment des expressions figées; à l'inverse, des schémas valides sont écartés car trop rares. Néanmoins, les expériences précédentes sur d'autres langues avec des techniques similaires l'ont bien montré, ce type d'analyse est efficace pour acquérir des données en largeur, pour un nouveau domaine ou une nouvelle langue. Comme nous le montrons dans les pages suivantes, il est alors possible d'acquérir de nouvelles données, de compléter des lexiques existants, voire de créer rapidement de nouveaux lexiques moyennant une analyse et une révision manuelle.

La tableau 6.4 met en paralèlle les chiffres de LEXSCHEM avec ceux d'autres lexiques de sous-catégorisaiton pour le français. Il est toutefois délicat de comparer ces chiffres étant donné les différences entre les lexiques, tant du point de vue de leur construction que des informations plus ou moins précises présentes dans les SSC. Par exemple, LEXSCHEM, DICOVALENCE, LGLEX et le LEFFF précisent quelles prépositions sont susceptibles de régir les compléments (voire dans certains cas quels types de compléments : il existe par exemple dans DICOVALENCE un paradigme locatif) alors que ce n'est pas le cas pour TREELEX qui ne différencie que 3 types de compléments prépositionnels : ceux régis par *à*, ceux régis par *de* et les autres. Il en est de même pour la finesse des ressources quant aux différents types de constructions pronominales (voir chapitre 3 et section 6.2.2). On observe

fieurs (voir chapitre 7). Ce point reste néanmoins un inconvénient au niveau de l'acquisition de SSC.

Lexique	LEXSCHEM	DICOVAL	LEFFF	LGLEX	TREELEX
Méthode	Auto	Manuelle	Mixte	Semi-auto	Semi-auto
Nb de verbes	3123	3738	6825	5694	2000
Nb d'entrées	7239	8313	?	16903	(4180)
Nb de SSC	88	?	?	?	180
Nb moyen de SSC par verbe	2,3	(2,22)	?	(2,97)	2,09

TABLE 6.4 – *Comparaison de* LEXSCHEM *avec d'autres lexiques de sous-catégorisaiton pour le français (les chiffres entre parenthèses ne sont pas des chiffres communiqués par les auteurs : ils résultent de nos calculs;* DICOVAL = DICOVALENCE*)*

toutefois une certaine proximité de LEXSCHEM avec DICOVALENCE en termes de nombre de verbes et d'entrées. Le nombre d'entrées de LGLEX est dû aux nombreuses exceptions (expressions semi-figées, verbes supports, etc) codées directement dans le lexique (voir section 3.1). On remarque aussi la faible couverture de TREELEX qui a été acquis à partir d'un corpus arboré d'un million de mots. Le relatif faible nombre de verbes de LEXSCHEM s'explique par les traitements visant à privilégier la qualité de la ressource par rapport à sa couverture (validation des verbes à la sortie de TREETAGGER, seuil sur le nombre d'occurrence des verbes, etc).

Format des entrées de LexSchem et extrait du lexique

Les informations disponibles pour chaque entrée de LEXSCHEM sont :

- l'identifiant de l'entrée;
- le verbe concerné par l'entrée;
- le schéma de sous-catégorisation concerné par l'entrée;
- le nombre d'occurrences de l'entrée dans le corpus;
- le nombre d'occurrences du verbe dans le corpus;
- le nombre de SSC pour ce verbe;
- la fréquence relative de l'entrée;
- les identifiants des séquences de l'analyse SYNTEX qui ont produit cette entrée;
- le nombre d'arguments du SSC;
- les lemmes têtes de chacun de ses arguments.

Nous présentons ici quelques exemples d'entrée du lexique[5] :

Les deux entrées du verbe *accompagner* dans LEXSCHEM sont :

```
*** Entrée 03739  (6135 occurences, fréquence relative : 75.4)
Verbe : accompagner (8142 occurrences, 2 ssc)
SSC : [SUJ:SN,OBJ:SN]
Séquences Syntex : 0192.anasynt!d226258p6_4!14, ...
Lemmes têtes de l'argument 0: il (477:0.08), groupe (36:0.01),
délégation (21:0.00), sauce (4:0.00), ...
Lemmes têtes de l'argument 1: le (1699:0.27), mouvement
(68:0.01), président (52:0.01), 0), beignet (1:0.00), ...
```

5. Le format de LEXSCHEM utilisé pour ces entrées est le format « normal » (voir section 6.4.2 pour un détail des différents formats disponibles pour LEXSCHEM).

```
*** Entrée 01629  (2007 occurences, fréquence relative : 24.6)
Verbe : accompagner (8142 occurrences, 2 ssc)
SSC : [SUJ:SN,REF:refl,DE-OBJ:SP<de+SN>]
Séquences Syntex : 0295.anasynt!d374460p4_1!63, ...
Lemmes têtes de l'argument 0: elle (212:0.10),
réduction (20:0.01), inactivité (1:0.00), ...
Lemmes têtes de l'argument 1: se (2007:1.00)
Lemmes têtes de l'argument 2: augmentation (31:0.02), réforme
(16:0.01), restructuration (10:0.00), chant (1:0.00), ...
```

Un certain nombre de constructions répertoriées dans les ressources présentées au chapitre 3 manquent dans LexSchem. C'est notamment le cas des schémas présents dans TreeLex et DicoValence qui correspondent aux constructions suivantes (extraites du corpus *LM10*) :

Pierre accompagne son fromage de vin.

Une fois cuit le pinon est servi chaud et s'accompagne avec le reste de la sauce tomate.

Ces SSC sont présents dans les sorties du constructeur de SSC candidats mais ils sont trop peu fréquents pour apparaître après filtrage :

[SUJ:SN, OBJ:SN, DE-OBJ:SP<de+SN>] (147 occurrences)
[SUJ:SN, REF:refl, P-OBJ:SP<avec+SN>] (4 occurrences)

Ces deux constructions apparaissent quasi-uniquement dans un contexte culinaire, peu fréquent dans le corpus LM10. Elles sont donc « écrasées » par les milliers d'occurrences des deux schémas principaux d'*accompagner*. Il faut nuancer l'absence de ces schémas par l'usage qui doit être fait des systèmes d'acquisition automatique : ils n'ont pas vocation à « concurrencer » les dictionnaires construits à la main sur la langue générale mais à adapter une ressource pour un corpus ou un domaine donné.

Voici les entrées du verbe *acheter* dans LexSchem :

```
*** Entrée 00615  (705 occurences, fréquence relative : 11.9)
Verbe : acheter (5914 occurences, 3 ssc)
SSC : [SUJ:SN,OBJ:SN,A-OBJ:SP<à+SN>]
Séquences Syntex : 0100.anasynt!d685792p5_3!72, ...
Lemmes têtes de l'argument 0: il (108:0.15), Chine (3:0.00),
...
Lemmes têtes de l'argument 1: le (53:0.07), action (14:0.02),
...
Lemmes têtes de l'argument 2: lui (85:0.12), Japon (17:0.02),
...

*** Entrée 06495  (1301 occurences, fréquence relative : 22.0)
Verbe : acheter (5914 occurences, 3 ssc)
SSC : [SUJ:SN]
Séquences Syntex : 1099.anasynt!d671009p2_3!20,...
Lemmes têtes de l'argument 0: musée (9:0.01), pays (7:0.01),
...

*** Entrée 03561  (3908 occurences, fréquence relative : 66.1)
```

```
Verbe : acheter (5914 occurences, 3 ssc)
SSC : [SUJ:SN,OBJ:SN]
Séquences Syntex : 0500.anasynt!d708862p3_7!2, ...
Lemmes têtes de l'argument 0: elle (171:0.04), pays (19:0.00),
...
Lemmes têtes de l'argument 1: le (272:0.07), livre (71:0.02),
...
```

Les SSC `[SUJ:SN,OBJ:SN,A-OBJ:SP<à+SN>]` (entrée `00615`) et `[SUJ:SN,OBJ:SN]` (entrée `03561`) sont valides. Voici quelques phrases du corpus à partir desquelles ces entrées ont été acquises :

*Le propriétaire à qui il l'**a achetée** était astrologue et lui a vendue à bas prix parce que les chiffres étaient favorables.* (`[SUJ:SN,OBJ:SN,A-OBJ:SP<à+SN>]`)

*Et les investisseurs qui les **ont** récemment **achetées** pourraient en profiter.* (`[SUJ:SN,OBJ:SN]`)

Toutefois, en ce qui concerne le SSC `[SUJ:SN,OBJ:SN,A-OBJ:SP<à+SN>]`, il peut également s'agir de phrases où le complément en *à* est un modifieur :

*A la sortie du musée , il **achète** un laissez-passer pour un an.*

L'entrée `06495` (schéma `[SUJ:SN]`) est quant à elle essentiellement due à des erreurs d'analyse syntaxique :

*Qu'en est-il, aujourd'hui, des intérêts des copropriétaires qui **ont acheté**, entre 1986 et 1989, des appartements non autorisés à la construction et détruits en 1989 par le FLNC ?*

Ces erreurs sont commentées plus en détail dans la section 6.2.2. Un extrait plus large de LexSchem est donné en annexe D.

6.1.3 Une interface web de consultation du lexique

Une interface de consultation a été développée en PHP/MySQL et mise en ligne[6]. Elle permet de consulter les entrées de LexSchem correspondant à un verbe ou à un SSC. On peut donc avoir un aperçu de tous les SSC correspondant à un verbe donné ou de tous les verbes qui acceptent un schéma donné. Pour chaque entrée consultée, l'interface présente son nombre d'occurrences, sa fréquence relative ainsi que 5 exemples extraits du corpus. Il est également possible de consulter les analyses Syntex correspondant aux exemples, ce qui permet d'avoir une meilleure visibilité sur l'origine des erreurs.

La première page de l'interface propose à l'utilisateur de choisir un verbe ou un SSC parmi deux listes déroulantes (voir figure 6.2). Selon le choix fait par l'utilisateur, l'interface propose la liste des SSC correspondant au verbe choisi (voir figure 6.3) ou la liste des verbes qui acceptent le SSC choisi. L'utilisateur est invité à choisir un SSC / un verbe pour consulter les informations correspondantes à ce choix (figures 6.4 et 6.5).

6. http://www-lipn.univ-paris13.fr/~messiant/lexschem.html

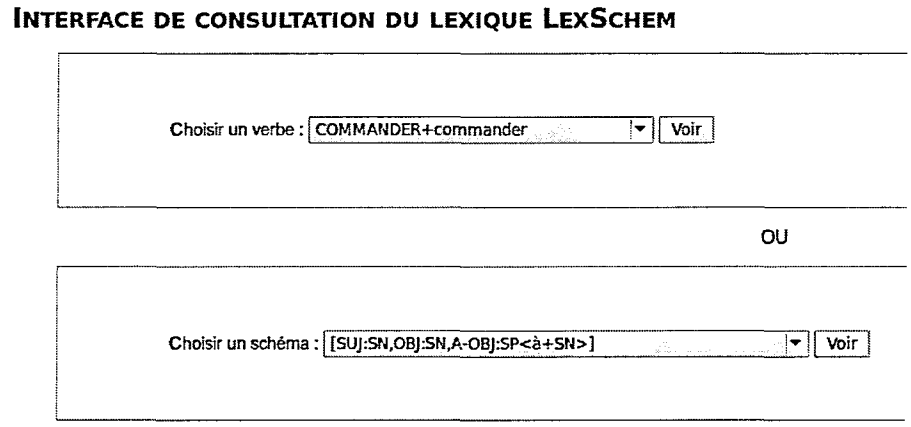


FIGURE 6.2 – *Page d'accueil de l'interface de consultation de* LEXSCHEM

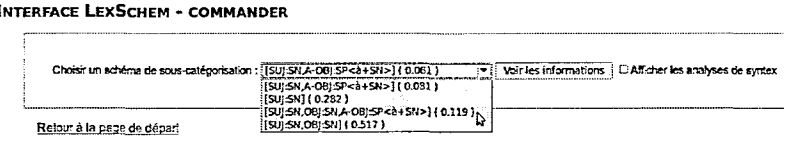


FIGURE 6.3 – *Choix du SSC pour le verbe commander dans l'interface*

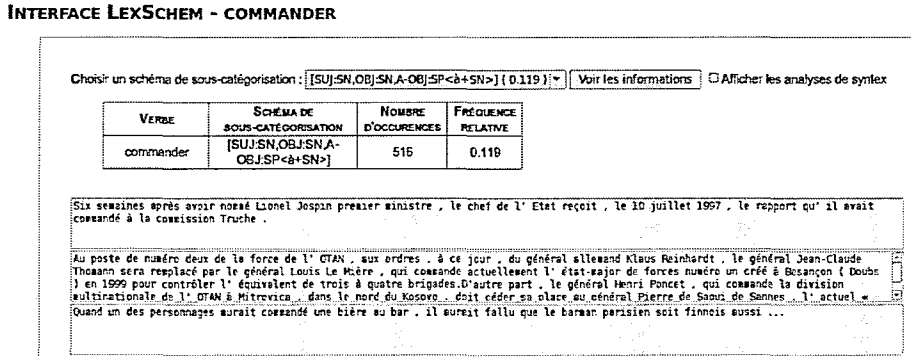


FIGURE 6.4 – *Résultats affichés dans l'interface pour l'entrée commander /* `[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]`

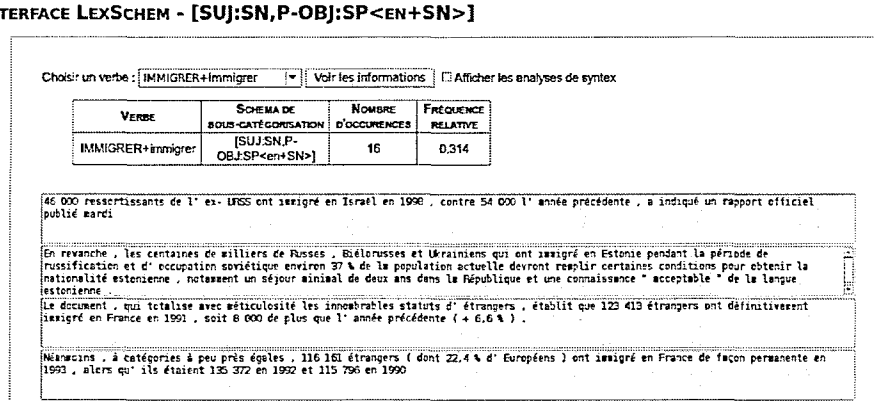


FIGURE 6.5 – *Résultats affichés dans l'interface pour l'entrée* `[SUJ:SN, P-OBJ:SP<en+SN>]` */ immigrer*

6.2 Évaluation de LexSchem

L'évaluation des ressources est un enjeu important qui permet de mettre en évidence les limites de ces ressources afin de pouvoir les améliorer. Les différentes méthodes d'évaluation ont été présentées à la section 4.1.

La qualité des ressources acquises par les systèmes automatiques fait l'objet de critiques. Premièrement, ces lexiques sont incomplets, c'est-à-dire que les informations qu'ils contiennent sont souvent moins riches que celles présentes dans les lexiques construits à la main. Deuxièmement, les informations acquises sont en partie erronées pour différentes raisons : imperfection de l'annotation morphosyntaxique ou de l'analyse syntaxique, erreurs lors du processus d'acquisition, difficulté de la distinction argument - modifieur. Enfin, il n'y a pas dans ces ressources d'entrées lexicales au sens défini à la section 2.1. En effet, il est difficile de distinguer différents sens pour un même verbe à l'aide de méthodes automatiques.

LexSchem est une ressource acquise automatiquement. Une évaluation du lexique est donc nécessaire afin d'identifier ses lacunes et de proposer des solutions quand cela est possible. L'évaluation permet également de donner un aperçu de la qualité de la ressource (et du système d'acquisition) aux utilisateurs potentiels.

Nous comparons tout d'abord LexSchem à d'autres lexiques du français avant d'examiner plus en détail la nature des informations qu'il contient et l'importance du type de corpus utilisé lors de l'acquisition. Nous proposons enfin une typologie des erreurs du système.

6.2.1 Évaluation quantitative : comparaison avec d'autres ressources

La manière la plus classique d'évaluer une ressource acquise automatiquement est de la comparer à d'autres ressources. Il s'agit de l'approche généralement adoptée, par exemple par Preiss et al. (2007).

Ressources utilisées et format commun

Nous sommes partis de deux lexiques de sous-catégorisation pour le français : TreeLex (Kupść et Abeillé, 2008b) et DicoValence (van den Eynde et Mertens, 2006), choisis pour leurs spécificités et leur disponibilité. Nous avons présenté ces ressources dans le chapitre 3 et nous avons vu qu'elles ne sont pas comparables, même si des similarités existent.

TreeLex est issu d'un corpus annoté (le corpus arboré de Paris 7) et DicoValence est le fruit d'un travail manuel (pour une présentation plus détaillée de ces deux lexiques, voir le chapitre 3). Les deux lexiques reposent donc, directement ou indirectement, sur un important travail de description linguistique préalable, ce qui permet de s'assurer de l'exactitude des schémas présents dans ces ressources (dans la grande majorité des cas). Cependant, cela ne garantit pas que ces lexiques soient complets, particulièrement pour TreeLex qui a été acquis à partir d'un corpus journalistique de taille limitée.

Il faut par ailleurs noter que DicoValence est structuré autour de la notion d'entrée lexicale : chaque entrée réflète une distinction de sens, comme dans un dictionnaire usuel (pour distinguer les sens différents d'un même

verbe, un champ fournit la traduction du verbe en anglais). Par contre, TREELEX fournit des listes de SSC pour un même lemme verbal « à plat », c'est-à-dire que les constructions sont directement associées à un lemme verbal et non à une unité lexicale reflétant les distinctions de sens. De ce point de vue, LEXSCHEM est proche de TREELEX ; l'absence d'entrées lexicales en tant que telles est certes une limite pour certaines applications mais n'est pas toujours gênante s'il s'agit de fournir des ressources à un analyseur syntaxique ou de fournir des données au linguiste qui se charge ensuite de les valider.

Pour des raisons différentes, ces deux ressources sont partielles : les concepteurs de DICOVALENCE se sont concentrés sur les verbes et les constructions essentielles du français, tandis que TREELEX n'inclut que ce qui est attesté dans le corpus arboré utilisé comme source. Il est donc intéressant de comparer la couverture relative de ces différentes ressources et l'apport possible de LEXSCHEM.

L'utilisation de DICOVALENCE pour évaluer LEXSCHEM nécessite une transformation de format délicate en raison des spécificités liées à l'Approche Pronominale. À des fins de comparaison et d'évaluation, Claire Gardent (Gardent et Lorenzo, 2010) a unifié un certain nombre de ressources dans un format pivot dans le cadre du projet TALC (Traitement automatique des langues et des connaissances)[7]. Le processus n'est évidemment pas sans poser problème : les ressources concernées ne sont pas fondées sur la même théorie, elles n'ont pas le même objectif et les informations n'y sont pas toujours codées de façon explicite. Nous nous sommes intéressés en particulier à DICOVALENCE-EASY qui est la transformation de DICOVALENCE dans ce format pivot[8]. Des difficultés de codage ont entraîné des erreurs : des constructions répertoriées dans DICOVALENCE sont absentes de DICOVALENCE-EASY (c'est notamment le cas des compléments infinitifs introduits par une préposition : `P-OBJ:SP<prep+SINF>` qui sont systématiquement marqués `OBJ:SINF` ou des propositions subordonnées). Un article récent de Piet Mertens (2010) confirme notre observation. Nous utiliserons donc DICOVALENCE-EASY à des fins de comparaison mais avec prudence. Dans le cas de DICOVALENCE, l'évaluation manuelle des résultats pour laquelle nous avons utilisé la ressource originale (voir section 6.2.2) est donc indissociable de la comparaison quantitative présentée ici.

Conversion de LexSchem

Pour effectuer la comparaison des lexiques, il a fallu trouver un format commun conservant au mieux les informations originales (mais, du fait des choix de descriptions variables d'un lexique à l'autre, la transformation ne peut être complètement neutre). Le format retenu est proche du format *Easy*, du fait que les versions de TREELEX et de DICOVALENCE utilisées sont déjà encodées dans ce format. Les compléments prépositionnels y sont réduits à trois catégories : `A-OBJ:SP<à+ SN|SINF>` (compléments en « *à* »), `DE-OBJ:SP<de+ SN|SINF>` (compléments en « de ») et `P-OBJ:SP<prep+ SN|SINF>` (les compléments gouvernés par toutes

7. Les lexiques au format Easy peuvent être trouvés à l'adresse suivante : `http://talc.loria.fr/`. Tous les lexiques mentionnés ici ont été consultés en ligne le 20 février 2010.

8. `http://talc.loria.fr/Dicovalence-Easy.html`

les autres prépositions). Ce processus de normalisation est nécessaire (par exemple, le détail des prépositions est absent dans TreeLex et n'est pas systématique dans DicoValence) mais provoque un appauvrissement certain des informations contenues dans le lexique. Toutefois, la phase de validation manuelle (présentée dans la section suivante) permet ensuite de vérifier que les schémas comprenant des P-OBJ concernent les bonnes prépositions.

Quelles mesures de comparaison des ressources ?

Nous avons expliqué plus haut (section 4.1) les problèmes liés à l'utilisation d'une ressource référence pour l'évaluation de ressources : le biais induit par la théorie sur laquelle repose la ressource, la variabilité du traitement de la distinction argument/modifieur dans les lexiques construits à la main, *etc.* sont autant de facteurs qui relativisent les résultats obtenus. Or, les ressources utilisées ici ont été acquises à partir de corpus arboré et/ou ont subi une transformation de format délicate.

Nous avons donc fait le choix de ne pas utiliser les mesures habituelles de précision, rappel et F-mesure, dans la mesure où elles impliquent que la ressource référence utilisée soit suffisamment complète et précise pour représenter « toute » la langue. De plus, ces chiffres masquent souvent l'hétérogénéité des erreurs : se confondent des entrées erronées en raison d'une erreur d'annotation ou d'analyse syntaxique, des entrées mal filtrées et des entrées contenant un modifieur (et on sait qu'il est parfois difficile, même pour un observateur humain de distinguer les arguments des modifieurs). Pour toutes ces raisons, la précision, le rappel et la F-mesure nous semblent trop grossières pour l'évaluation d'une ressource comme LexSchem.

Lorsqu'on acquiert une nouvelle ressource, on cherche à mesurer à la fois son recouvrement avec les ressources existantes et la nouveauté qu'elle apporte par rapport à ces ressources (cette nouveauté peut être notamment due au corpus utilisé pour l'acquisition ou à des lacunes dans les ressources existantes). La « nouveauté » est le nombre d'entrées acquises automatiquement et qui peuvent compléter une ressource existante. Habituellement, cette nouveauté est traitée comme un déficit de précision de la ressource évaluée. Or, si les entrées nouvelles contiennent une part d'entrées incorrectes, elles contiennent également des entrées absentes de la ressource référence mais qui correspondent à un usage dans la langue (autrement dit, ce n'est pas parce qu'une entrée n'était pas dans la ressource de référence qu'elle est fausse ; c'est précisément cette nouveauté qui fait tout l'intérêt des méthodes automatiques). Il s'agira dans un deuxième temps de mesurer la qualité des schémas, qu'ils soient nouveaux ou déjà mentionnés dans les ressources de référence par une analyse qualitative (voir section 6.2.2). Cette approche permet de mettre en évidence le réel apport de notre approche par rapport à l'état de l'art.

Pour mesurer le « recouvrement » et la « nouveauté » de LexSchem par rapport à une autre ressource, nous avons donc comptabilisé le nombre d'entrées communes aux deux ressources, la proportion de ces entrées par rapport au nombre total d'entrées dans le lexique de comparaison (recouvrement) et le nombre d'entrées présentes dans LexSchem et non dans la ressource de comparaison (indice de la nouveauté).

	TreeLex	DicoValence-EASy
Nb total d'entrées dans la ressource	3570	4282
Nb d'entrées communes avec LexSchem	2181	2563
Recouvrement	61,1%	59,9%
Nb d'entrées nouvelles dans LexSchem	1440	1058

Table 6.5 – *Nombre d'entrées communes et nouvelles dans* LexSchem *par rapport aux deux ressources de référence (pour les 1583 verbes communs aux 3 ressources)*

$$recouvrement = \frac{card(eval \cap comp)}{card(comp)}$$

(où *eval* est l'ensemble des entrées de la ressource évaluée et *comp*, celui des entrées de la ressource de comparaison)

Si les couples verbe–SSC qui figurent à la fois dans LexSchem et dans les ressources de référence sont généralement bons, il est plus délicat de valider ou d'invalider, hors contexte, les nouveaux SSC proposés par LexSchem pour chaque verbe. Un travail manuel de validation est alors nécessaire. Les entrées absentes de LexSchem mais présentes dans la ressource de comparaison posent un problème encore plus délicat : l'entrée n'a-t-elle pas été repérée à cause d'erreurs d'analyse ou est-elle juste absente du journal *Le Monde* (comme c'est le cas pour les constructions du verbe *accompagner* présentées plus haut) ? Au-delà de quelques investigations manuelles pour des verbes particuliers, il est difficile d'être catégorique quant au silence de la ressource.

Résultats et discussion

Les SSC des 1583 verbes communs aux 3 ressources ont été comparés. Le tableau 6.5 montre les résultats obtenus. Précisons toutefois que le problème de transformation de format de DicoValence vers DicoValence-EASy provoque probablement une surrestimation de la nouveauté de LexSchem par rapport à DicoValence-EASy. Ce problème de transformation de format peut aussi avoir un effet, plutôt positif cette fois, sur le recouvrement entre LexSchem et DicoValence. L'investigation manuelle est donc très importante pour étudier ce biais.

On ne remarque pas de différence significative au niveau du recouvrement entre les deux ressources de référence, TreeLex et DicoValence-EASy. La seule différence remarquable concerne les nouveaux SSC, plus nombreux quand on compare LexSchem avec TreeLex qu'avec DicoValence-EASy. Ceci s'explique probablement par le plus grand nombre de SSC dans DicoValence (donc la meilleure couverture de cette ressource ; TreeLex a été acquis sur un corpus de taille moyenne — 1 million de mots — qui ne contient qu'un sous ensemble des SSC caractéristiques du français). Notons également que la nouveauté estimée pour DicoValence-EASy est probablement surestimée étant donné les problèmes d'alignement évoqués plus haut.

Une part non négligeable des SSC des ressources de référence n'est pas retrouvée par notre méthode. Ceci est dû à la stratégie de filtrage qui élimine les SSC les moins fréquents. Il s'agit d'un biais assez fréquent des techniques statistiques, qu'il est difficile de corriger au niveau du filtrage lui-même : des techniques de filtrage plus sophistiquées n'ont pas montré d'améliora-

tion significative (Korhonen et al., 2000). On s'aperçoit en revanche lors de l'analyse manuelle qu'un nombre non négligeable de SSC manquants pourrait être inféré à partir des SSC effectivement repérés : une bonne partie des SSC manquants sont en fait des formes réduites de SSC complexes (du fait du caractère optionnel dans les réalisations de surface de la plupart des arguments). Ainsi, le système a inféré que le verbe *donner* pouvait être employé avec deux compléments (*Les anticolonialistes de ce bord-là donnaient une dimension morale à leur choix*) ou un complément d'objet direct seul (*Il donne l'alerte*). Le système peut assez sûrement proposer dans ce type de cas la construction avec un complément d'objet indirect seul (*Il faut donner au denier du culte*), surtout si quelques exemples ont été trouvés en corpus. Il est donc possible d'améliorer la couverture en introduisant un processus de ce type en post-traitement afin de réduire le nombre de SSC manquants. Une étude préliminaire de ce type est présentée à la section 6.4.3 On arrive alors à une couverture plus satisfaisante de LexSchem par rapport aux ressources existantes.

Le tableau 6.5 montre également l'apport possible de LexSchem par rapport aux deux ressources de référence (sans prise en compte du post-traitement que nous venons d'évoquer). Les résultats doivent être complétés par une analyse manuelle des SSC obtenus afin de vérifier la pertinence des nouveaux schémas trouvés et la qualité des ressources de référence par rapport à notre tâche. Nous avons effectué, en collaboration, une évaluation manuelle sur 150 verbes (voir section 6.2.2). Cette évaluation a révélé que 109 nouvelles entrées valides pouvaient être ajoutés à TreeLex et 33 à DicoValence-EASy[9]. Plus de la moitié des entrées nouvelles proposées pour ces 150 verbes ont été validées par les annotateurs (alors que ces entrées auraient été purement et simplement considérées comme « fausses » par une évaluation classique par la mesure de la précision et du rappel). Ceci montre selon nous la capacité de notre système (et plus généralement des méthodes automatiques) à assister le linguiste lors de l'élaboration d'un lexique ou lorsqu'il souhaite compléter des ressources existantes.

6.2.2 Évaluation qualitative

Une analyse qualitative des ressources obtenues par acquisition automatique est indispensable. En effet, la comparaison de LexSchem à d'autres ressources ne donne qu'un aperçu partiel de la qualité du lexique de par le caractère complexe de la nouveauté : un schéma « nouveau » dans LexSchem peut être dû à une erreur ou au contraire être un schéma absent des autres ressources mais valide pour la tâche visée (sur ce problème voir les sections 4.1 et 6.2.1). Nous avons donc effectué une analyse manuelle de la couverture et de la nouveauté obtenues lors de la comparaison de LexSchem avec TreeLex et DicoValence-EASy. Nous nous penchons ensuite sur le cas des constructions pronominales. Enfin, les analyses manuelles de LexSchem

9. Il faut toutefois noter que c'est à dessein que DicoValence et TreeLex ne sont pas exhaustifs : les concepteurs de DicoValence ont par exemple volontairement limité leur lexique aux 3 500 verbes les plus fréquents du français. Il n'empêche que les applications de traitement des langues exigent des dictionnaires aussi complets et précis que possible et, dans ce cadre, les méthodes (semi-)automatiques peuvent se révéler précieuses.

nous ont permis de mettre en évidence une typologie des erreurs rencontrées dans notre ressource.

Analyse manuelle de la couverture et de la nouveauté

Les entrées de 150 verbes variés (choisis en fonction de leur fréquence, de leur généricité et du nombre de SSC qui leur est associé) ont été évalués par deux annotateurs indépendamment afin de mieux mesurer la qualité de LEXSCHEM [10].

Nous avons généré un fichier XML comprenant les SSC pour 150 verbes parmi les 1583 verbes utilisés lors de la comparaison de LEXSCHEM avec les deux autres lexiques (section 6.2.1). Pour chacun de ces verbes, nous avons regroupé les entrées apparaissant dans LEXSCHEM, TREELEX ou DICOVALENCE-EASY. Chaque entrée est caractérisée par sa présence ou son absence dans LEXSCHEM d'une part, dans TREELEX ou DICOVALENCE-EASY d'autre part, ainsi que par 5 exemples tirés du corpus *Le Monde* (dans le cas où l'entrée est présente dans LEXSCHEM). De plus, afin d'éviter les erreurs dues à la transformation de format, un retour à DICOVALENCE est réalisé de manière systématique lorsque certaines entrées évidentes sont absentes de DICOVALENCE-EASY ou que des entrées fausses sont présentes dans DICOVALENCE-EASY.

L'évaluation des entrées se fait grâce aux balises `<hum>` et `<Commentaires>`. La balise `<hum>` permet à l'annotateur de caractériser chaque entrée selon un code défini en amont. Ce code permet de caractériser l'entrée de LEXSCHEM par rapport aux autres ressources et de retrouver facilement des exemples correspondants à chacun des cas.

Un guide d'annotation a été rédigé pour garantir la cohérence de l'évaluation. Un code a été défini afin de caractériser les entrées du fichier dans la balise `<hum>`.

- les entrées de LEXSCHEM jugées inexactes par l'annotateur ;
- les entrées de LEXSCHEM jugées exactes ;
- les entrées présentes dans TREELEX ou DICOVALENCE mais absentes de LEXSCHEM qui semblent correctes, dans le sens où elles réflètent bien un comportement de sous-catégorisation (*i.e.* pour lesquelles l'annotateur trouve facilement un exemple) ;
- les entrées présentes dans TREELEX ou DICOVALENCE mais absentes de LEXSCHEM qui semblent incorrectes ;
- les entrées présentes dans TREELEX ou DICOVALENCE mais absentes de LEXSCHEM pour lesquelles l'annotateur n'est pas capable de dire hors contexte si elles sont correctes ou non.

Une balise de commentaires permet à l'annotateur de faire des remarques sur l'entrée concernée : est-ce que l'entrée correspond bien aux exemples ? A quoi sont dues les erreurs ? etc.

Un extrait de l'entrée du fichier d'évaluation correspondant au schéma `[SUJ:SN,OBJ:SN,A-OBJ:SP<à+SN>]` pour le verbe « *commander* » est donné en figure 6.6

10. Il s'agit de Thierry Poibeau et de nous-même. Un guide d'annotation a toutefois été rédigé pour garantir la cohérence de l'évaluation.

```
<verbe lemme="commander">
    <ssc forme="[SUJ:SN,OBJ:SN,A-OBJ:SP<à+SN>]"
         caracteristiques="present-lexschem;
   absent-treelex;
   present-dicovalence;">
        <hum></hum>
        <Commentaires></Commentaires>
        <exemple>
            La Shochiku lui commande Yokai Hanta Hiruko qui a
            droit , lui , à une sortie commerciale .
        </exemple>
```

FIGURE 6.6 – *Extrait du fichier d'évaluation*

	Annotateur A		Annotateur B	
Entrées de LEXSCHEM :				
incorrectes	39	6,5%	49	8,4%
correctes	311	53,5%	301	51,8%
Entrées absentes de LEXSCHEM :				
correctes	131	22,5%	137	23,6%
douteuses ou incorrectes	100	17,4%	94	16,2%
Apport à TREELEX	116 nouvelles entrées		109 nouvelles entrées	
Apport à DICOVALENCE	42 nouvelles entrées		33 nouvelles entrées	

TABLE 6.6 – *Données issues de l'analyse manuelle de 150 verbes de* LEXSCHEM

Au total, 581 entrées ont été annotées pour 150 verbes[11]. Les résultats sont présentés dans le tableau 6.6.

Pour les formes absentes de LEXSCHEM, l'annotation prévoyait initialement deux catégories « douteux » d'une part et « incorrect » d'autre part. Nous avons eu beaucoup de mal à juger de la pertinence de certaines entrées hors contexte, ce qui a entraîné de fortes variations dans l'évaluation entre ces deux catégories qu'il est en fait difficile de distinguer. Nous avons donc choisi de les fusionner ici.

Cette analyse montre qu'il manque dans LEXSCHEM environ 40 % des entrées présentes dans les autres lexiques. Ceci est intéressant dans la mesure où une part non négligeable de ces entrées sont jugées incorrectes ou douteuses par les évaluateurs humains. Ceci montre que même des lexiques construits manuellement ou à partir de données fortement validées à la main ne sont pas parfaits : ils peuvent contenir des erreurs ou des constructions si rares qu'elles sont jugées douteuses par les évaluateurs. Le fait que ni DICOVALENCE ni TREELEX ne comporte d'informations sur la productivité des constructions indiquées ne permet pas de relativiser le poids de constructions peut-être rares ou spécialisées.

Par ailleurs, ces résultats montrent l'intérêt de l'approche automatique pour compléter des ressources existantes. En effet, l'évaluation a révélé plus de 100 entrées valides absentes de TREELEX et environ 40 dans le cas de DICOVALENCE. Par exemple, le SSC `[SUJ:SN,REF:refl,DE-OBJ:SP<de+SN>]`

11. Chacune de ces entrées provient d'au moins une des trois ressources (mais certaines d'entre elles sont absentes de LEXSCHEM), ce qui explique le nombre important d'entrées par rapport à la moyenne de 2,2 entrées par verbe évoquée plus haut.

n'est pas présent dans *TreeLex* pour le verbe `détacher`. Pourtant, cette entrée est présente dans LexSchem et est vérifiée par des exemples tirés du corpus comme « *Enfin , l' Albanie se détacha à son tour de la Chine , devenue par la force des choses son alliée .* ».

Cet expérience montre aussi que DicoValence est un lexique assez complet : LexSchem lui apporte peu d'entrées nouvelles et correctes (entre 30 et 45). Parmi ces entrées, on trouve le SSC `SUJ:SN,A-OBJ:SP<à+SN>` pour le verbe *virer* qui correspond bien à des usages corrects dans *Le Monde* :

une éthique qui vire à l'impuissance
il a viré au scepticisme
s'il vire au bleu marine, c'est la catastrophe

L'absence de cet emploi assez courant montre l'incomplétude des ressources construites manuellement, même lorsque celles-ci ont une bonne couverture. Parmi les entrées présentes dans LexSchem mais absentes des deux autres ressources, on trouve notamment la construction `[SUJ:SN,OBJ:SN,A-OBJ:SP<à+SN>]` pour le verbe *imprimer*. Si on peut penser au premier abord que ce schéma correspond à des expressions figées (comme dans l'exemple *Bonitzer imprime à son étude la marque de ses propres thèmes fétiches*), ce n'est pas toujours le cas dans les exemples extraits du corpus :

Maître exigeant, il imprima sa passion à des cuisiniers aussi différents que Pacaud, Dutournier ou Massonnet à Poitiers.

Les lois de la mécanique céleste prévoient que, si un corps suffisamment massif tourne autour d'une étoile, il lui imprime un léger mouvement oscillatoire.

Des constructions courantes sont ainsi identifiées et montrent l'apport de l'acquisition automatique par rapport aux dictionnaires existants.

Le cas des constructions pronominales

Les constructions pronominales sont intéressantes parce qu'elles posent des problèmes de codage complexes, qui sont résolus de manière différente suivant les lexiques considérés. Il s'agit surtout d'un phénomène massif en français, qui concerne la plupart des lemmes verbaux.

Prenons le cas du verbe *confondre*. Pour ce verbe, LexSchem propose les SSC donnés dans le tableau 6.7 (L'étiquette `REF:refl` est utilisée pour toutes les constructions pronominales, indépendamment de la valeur sémantique du pronom).

Les constructions pronominales sont souvent couvertes de manière partielle dans les différents lexiques syntaxiques du français (voir section 3), y compris dans des lexiques par ailleurs quasi exhaustifs. Pour le verbe *confondre*, DicoValence est le plus précis dans la mesure où le lexique encode non seulement les formes pronominales mais précise en outre la valeur sémantique des arguments [12]. Le Lexique grammaire (et les versions dérivées comme LGLex) ne mentionne pas les constructions pronominales mais précise les réalisations possibles des arguments [13]. TreeLex ne mentionne qu'une seule construction, non pronominale (`[SUJ:SN,OBJ:SN]`).

12. Entrées 18280, 18290, 18300, 18305 et 18310 de DicoValence.
13. Entrées 4_114 et 32H_153 de LGLex.

SSC	Exemple
[SUJ:SN,OBJ:SN]	*Ces gentils faux bénévoles confondaient la chose et son slogan.*
[SUJ:SN,REF:refl]	*Si les oiseaux avaient la télé, il y a longtemps que leurs chants se confondraient.*
[SUJ:SN,OBJ:SN, P-OBJ:SP<avec+SN>]	*Un chasseur a été tué par erreur, par un compagnon qui l' a confondu avec un sanglier.*
[SUJ:SN,REF:refl, P-OBJ:SP<avec+SN>]	*Le roi est le patron du makhzen, mais il ne se confond pas avec lui.*

TABLE 6.7 – *SSC correspondants au verbe confondre dans* LEXSCHEM

Dans ce cas précis, les données de LEXSCHEM sont intéressantes et pourraient compléter certains des lexiques examinés ci-dessus. Il va de soi que pour d'autres cas moins favorables, ce serait les autres ressources qui seraient les mieux armées. Un travail reste donc nécessaire pour combiner les ressources et les diverses informations disponibles. Ceci est particulièrement vrai des formes pronominales : il est rapidement nécessaire de quitter le cadre purement syntaxique pour déterminer leur valeur sémantique (réfléchi, réciproque, passif, "pseudo-se", *etc*). Seul un travail manuel peut fournir des informations fines à ce niveau ; une stratégie possible est la fusion de sources de connaissances complémentaires : c'est par exemple la stratégie présentée pour compléter le Lefff sur cet aspect (Sagot et Tolone, 2009).

Typologie des erreurs observées dans LexSchem

L'évaluation manuelle décrite ci-dessus a également permis de dégager les causes d'erreurs les plus fréquentes dans LEXSCHEM et d'envisager des solutions dans le système d'acquisition pour éviter ces erreurs.

Nous avons choisi d'utiliser le couple TREETAGGER-SYNTEX pour l'analyse syntaxique car ces outils étaient les plus performants lorsque cette étude a été lancée, au moins sur les corpus de presse (cf. section 5.2). Ces outils sont cependant à l'origine de certaines erreurs d'analyse. Certaines erreurs dues au TREETAGGER ont été corrigées par le module de validation des verbes (voir section 5.3). On a ainsi pu diminuer drastiquement la proportion de noms étiquetés comme verbe (5 574 verbes douteux rejetés). Par exemple, le nom du pilote de formule un *Senna* n'est plus reconnu comme une forme du verbe *senner* et « *le somnanbule* » comme une forme du verbe *somnanbuler*.

Pour l'analyse syntaxique, SYNTEX a une stratégie prudente : si l'analyseur ne trouve pas d'indice suffisamment fiable pour le rattachement d'un complément, celui-ci peut être laissé "libre", c'est-à-dire qu'il flotte et qu'il n'est rattaché à aucun élément de la phrase (Bourigault et al., 2005). C'est notamment le cas dans certaines phrases comprenant des insertions, comme par exemple « *Il commande ensuite, sur Internet, des pièces détachées, qui donneront une arme parfaitement inutilisable.* » qui produit le schéma (erroné) [SUJ:SN] (soit verbe intransitif). Il arrive également que des pronoms ne soient pas rattachés au verbe par l'analyseur. Par exemple, la phrase *Tu couches ou je te vire* produit le schéma [SUJ:SN] (intransitif) pour le verbe *virer*. Nous avons partiellement répondu à ce problème en mettant un seuil

plus élevé pour le SSC [SUJ:SN], fréquemment produit à cause de ces erreurs d'analyse. Bien évidemment, cette stratégie n'est pas toujours suffisante.

Un dernier ensemble d'erreurs est lié au système d'acquisition lui même. Certains SSC sont incorrects parce qu'ils contiennent des modifieurs. Par exemple, le schéma [SUJ:SN_P-OBJ:SP<dans+SN>] est très présent en corpus pour le verbe dormir mais le complément introduit par la préposition *dans* correspond toujours à un complément circonstanciel de lieu : « *Il dort dans son lit.* ». La distinction argument–modifieur reste donc difficile quand on se fonde uniquement sur des indices de surface. Notons toutefois que ce type d'erreurs (et plus généralement la présence régulière de certains types de modifieurs) est utile pour le calcul de classes syntaxico-sémantique de verbes (voir chapitre 7) voire pour le rattachement de compléments fréquents au verbe par des analyseurs syntaxiques.

6.2.3 Conclusion sur l'évaluation

L'évaluation tant quantitative que qualitative de LexSchem a montré ce que l'acquisition automatique d'informations de sous-catégorisation verbale pouvait apporter aux ressources existantes. Elle a également montré les limites des ressources acquises automatiquement lorsqu'on les compare à des ressources construites ou validées « à la main » et portant sur la langue générale. La prochaine section explore les possibilités offertes par les systèmes d'acquisition automatique lorsque l'objectif est de développer une ressource pour une langue plus « spécialisée ».

6.3 Adaptabilité des méthodes d'acquisition automatique

L'un des principaux avantages des méthodes d'acquisition automatique de ressources lexicales est leur « adaptabilité », c'est-à-dire leur faculté à acquérir des ressources différentes en fonction des réglages ou des corpus utilisés. Nous montrons ici que la méthode d'acquisition présentée dans le chapitre 5 peut-être utilisée pour acquérir des ressources sur un corpus « spécialisé ». Nous abordons également la possibilité d'adapter notre méthode pour acquérir des SSC de prédicats non verbaux.

6.3.1 Acquisition de SSC de verbes pour une langue de spécialité

Les ressources « spécialisées » sont importantes pour améliorer la qualité des systèmes de traitement de la langue. En effet, les applications de traitement automatique des langues reposent de plus en plus sur des ressources lexicales. Or, les ressources disponibles concernent le plus souvent la langue « générale » et les ressources « spécialisées » sont pour l'instant peu disponibles. Il est difficile de définir les concepts de « langue générale » et de « langue de spécialité ». Ces notions sont discutées et leurs frontières sont floues : à partir de quand une langue est suffisamment spécifique pour être considérée comme spécialisée et méritant un traitement particulier ? À défaut d'être capables de définir des règles pour répondre à cette question, nous pensons que certaines langues professionnelles ou techniques sont très

spécifiques et peuvent être étudiées en tant que « langues de spécialité ». C'est notamment le cas des domaines juridiques et médicaux, souvent étudiés dans les laboratoires de linguistique. À titre d'exemple, nous avons choisi le corpus LM10 pour son hétérogénéité mais on ne peut pas considérer que ce corpus journalistique couvre toute la langue générale (comme nous l'avons montré à la section 6.1.2 avec l'absence de constructions relatives à l'utilisation du verbe *accompagner* dans un contexte culinaire). La construction manuelle de nouvelles ressources spécialisées est coûteuse et l'acquisition automatique d'une base de travail qui peut être ensuite validée et complétée « à la main » semble une piste intéressante.

Nous avons donc utilisé un nouveau corpus afin de déterminer la capacité d'ASSCi à acquérir de nouveaux SSC en fonction du corpus considéré. Pour ce faire, nous avons choisi de prendre comme source la partie française du corpus EUROPARL [14] (Koehn, 2005). Il s'agit d'un corpus parallèle librement disponible sur Internet, constitué des actes du Parlement Européen entre mars 1996 et septembre 2003 [15]. Même s'il ne s'agit pas d'un corpus spécialisé au sens propre du terme, on peut s'attendre au sein du corpus EUROPARL a avoir affaire à des SSC particuliers liés à la nature et au genre du corpus considéré, c'est-à-dire des textes législatifs européens.

Pour acquérir des SSC à partir du corpus EUROPARL, nous avons utilisé le système ASSCi présenté dans le chapitre 5 avec les réglages par défaut (notamment en ce qui concerne les seuils). L'objectif de cette étude est de valider notre hypothèse selon laquelle l'acquisition automatique de SSC sur un corpus spécialisé donne des résultats sensiblement différents de ceux obtenus lors de notre première expérience 6.1.

Le tableau 6.8 présente les chiffres obtenus pour le lexique acquis à partir du corpus EUROPARL (LS-EUROPARL) à ceux extraits du corpus *LM10* (LS-LM10, section 6.1). Le repérage automatique des nouveaux SSC est simple, dans la mesure où il suffit de les extraire par comparaison entre les deux ressources. 1 171 verbes sont présents à la fois dans LEXSCHEM-LM10 et dans LEXSCHEM-EUROPARL : tous les verbes présents dans EUROPARL sont également présents dans LM10 (ce qui s'explique par la différence de taille entre les deux corpus et le fait que *Le Monde* relate régulièrement des informations du domaine législatif). 2 445 entrées (couples verbe-SSC) sont présentes dans les deux ressources tandis que 4 794 couples verbe-schéma sont spécifiques à LEXSCHEM-LM10 et 258 à LEXSCHEM-EUROPARL. Ces chiffres montrent que les données présentes dans le lexique sont fortement dépendantes du corpus utilisé. Une analyse manuelle des données est nécessaire pour évaluer la spécificité des entrées issues du corpus EUROPARL par rapport à celles issues de LM10.

Une étude qualitative a donc ensuite été menée. Nous avons examiné les SSC spécifiques à LEXSCHEM-EUROPARL obtenus à partir de l'analyse automatique décrite auparavant. Parmi les SSC nouveaux repérés, on constate

14. `http://www.statmt.org/europarl/`.

15. Le corpus inclut 11 langues européennes : français, italien, espagnol, portugais, anglais, néerlandais, allemand, danois, suédois, grec et finnois. Chaque langue comprend environ 1 million de phrases, qui contiennent de l'ordre de 28 millions de mots ; nous n'utilisons ici que la partie française du corpus.

	Nb de verbes	Nb de SSC	Nb d'entrées
LS-LM10	3123	88	7239
LS-LM10 uniquement	1952	19	4794
LS-LM10 et LS-EuroParl	1171	69	2445
LS-EuroParl uniquement	0	5	258
LS-EuroParl	1171	74	2703

TABLE 6.8 – *Comparaison de* LEXSCHEM-EUROPARL *avec* LEXSCHEM-LM10

un grand nombre de constructions qui, bien qu'elles ressortissent à la langue générale, n'en sont pas moins remarquables dans le corpus EUROPARL. Ainsi, *enchaîner* se construit fréquemment avec les prépositions *avec* ou *sur*. Ceci est bien évidemment dû au contexte du parlement européen fait de longues séries de débats et de discussions (*J'enchaînerai immédiatement sur le thème évoqué par M. Ilgenfritz , celui des régions frontalières ; J'enchaînerai ensuite avec le rapport Purvis*). Dans le corpus LM10, *enchaîner* est le plus souvent utilisé au style direct (« *Il n' y a rien de plus excitant que de faire une pièce* », *enchaîne Martin Gousset.* ou *Elles enchaînent : « Non à l'amnistie »*), alors que cet emploi est quasi inexistant dans le corpus EUROPARL. Dans les deux corpus, le sens concret de *(s')enchaîner* est quasi absent.

On constate également l'apparition de nouveaux SSC et de familles sémantiques associées au contexte particulier du corpus EUROPARL. Ainsi, les constructions de *voter* avec les prépositions *pour, contre* ou la locution *en faveur de* sont sans commune mesure avec ce que l'on obtient à partir de LM10 (où seul *pour* est assez présent pour être conservé malgré le filtrage). On a, dans le même ordre d'idées, le verbe *se prononcer* qui peut se construire avec *sur* mais surtout *en faveur de, mettre à disposition* et *donner mandat* qui se construisent avec *pour, légiférer* et *s'abstenir* avec *sur, etc.* Tous ces verbes sont remarquables du domaine considéré et esquissent les éléments d'un sous-langage législatif.

Cette première expérience montre bien l'adaptabilité du système et sa capacité à acquérir des informations spécifiques selon le corpus utilisé. Néanmoins, dans le cas du domaine législatif, les SSC obtenus concernent des verbes spécifiques qui sont parfois utilisés dans le corpus journalistique LM10 qui relate régulièrement certains débats parlementaires.

L'acquisition de ressources à partir d'un corpus dont la « spécialisation » est encore plus marquée (un corpus médical par exemple) est donc tout à fait envisageable. La principale difficulté réside dans l'adaptation de TREETAGGER et de SYNTEX au domaine médical (par exemple pour qu'ils puissent reconnaître les noms de molécules). Une part de travail manuel est donc nécessaire. Toutefois, ce travail reste bien moins coûteux que la construction d'un lexique de sous-catégorisation « à la main ».

Notons également qu'il est possible d'utiliser des réglages plus « permissifs » (c'est-à-dire des seuils moins élevés) lors de l'acquisition afin de faire émerger un plus grand nombre de constructions. Le fait de privilégier le rappel sur la précision est une bonne solution lorsque l'on est assuré d'une validation manuelle de la ressource *a posteriori*. C'est pourquoi nous avons

rendu disponibles une version non filtrée de LexSchem et un programme de filtrage autonome [16] (voir section 6.4.1).

En tous cas, l'acquisition automatique de SSC peut fournir une base de travail intéressante pour les linguistes cherchant à bâtir des ressources spécialisées en leur permettant de se concentrer sur les cas difficiles ou peu fréquents.

6.3.2 Acquisition de SSC de prédicats non verbaux

Les méthodes d'acquisition automatique de ressources lexicales permettent également d'acquérir facilement des ressources pour d'autres types de prédicats. Par exemple, Judita Preiss et al. (2007) ont utilisé le système d'acquisition de SSC de Cambridge pour acquérir les SSC des adjectifs et des noms prédicatifs en anglais.

À l'image des études concernant les domaines de spécialité, la faible disponibilité de ressources de sous-catégorisation pour les prédicats non-verbaux justifie les expériences de ce type.

L'adaptation d'un système comme ASSCi à un autre type de prédicats est plus coûteuse que son utilisation sur un corpus de spécialité. Il reste néanmoins beaucoup plus simple et rapide d'acquérir automatiquement des SSC pour les noms prédicatifs ou les adjectifs que de développer une ressource « à la main » à partir de zéro (si on considère les travaux antérieurs de construction de dictionnaires). La qualité de la ressource acquise automatiquement dépend ici aussi de la taille du corpus utilisé et du temps et de l'attention attribués aux réglages du système et à la validation de la ressource.

Nous avons réalisé une première expérience d'acquisition automatique de SSC pour des noms prédicatifs en effectuant quelques modifications dans notre système : on cherche à repérer les constructions dans le corpus qui correspondent aux occurrences des noms prédicatifs. L'architecture du système ainsi que l'implémentation des modules a été très peu modifiée (nous avons par exemple utilisé la même liste d'étiquettes pour les compléments) pour cette expérience préliminaire qui avait pour objectif d'explorer la faisabilité de l'acquisition automatique de SSC de noms prédicatifs en français. Cette expérience faisait suite à la demande de linguistes dont le thème de recherche est la nominalisation (par exemple, le verbe *annuler* se nominalise en *annulation*). Le corpus que nous avons utilisé pour cette expérience est le corpus journalistique LM10 (voir section 6.1.1).

La principale difficulté rencontrée a été le repérage des noms prédicatifs dans le corpus : aucune information de surface ne nous permet de différencier *le camion de Jean* de *la construction de l'aqueduc*. La solution la plus simple à ce problème est de fournir une liste de noms « ciblés » au système. Nous avons décidé de traiter les noms répertoriés dans la ressource VerbAction qui est un lexique de noms d'actions morphologiquement apparentés à des verbes [17]. Par cette méthode, nous avons acquis les SSC de 1 436 noms prédicatifs.

16. http://www-lipn.univ-paris13.fr/~messiant/lexschem.html
17. http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=hathout/verbaction/main.html

Bien sûr, les SSC de noms prédicatifs sont très différents des SSC de verbes : les noms prédicatifs ne prennent ni sujet ni objet et le schéma intransitif (sans complément) n'a pas de sens (tous les noms peuvent apparaître sans complément). Tous les schémas acquis par notre méthode sont des combinaisons d'un ou de plusieurs syntagmes prépositionnels. Par exemple, le schéma SP<de+SN> apparaît pour de très nombreux noms dont *construction*, *consultation* et *perturbation*. Ce schéma est issu d'une propriété bien connue de la nominalisation :

Les ouvriers construisent le pont. → *La construction du pont*

La ressource acquise lors de cette première expérience n'a pas fait l'objet d'une évaluation approfondie. Il est néanmoins possible de reprendre cette étude pour acquérir des SSC de noms et/ou d'ajectifs prédidatifs à l'aide d'ASSCi. Ces informations pourront constituer une base intéressante pour la création d'un lexique de sous-catégorisation de ces prédicats pour le français [18].

6.4 Des outils pour le paramétrage des lexiques

Les informations utiles dans les lexiques de sous-catégorisation peuvent varier selon les besoins ou les applications visées : les applications de TAL reposent en majorité sur un lexique le plus exact possible mais le calcul de classes de comportement lexico-sémantiques peut profiter de la version non filtrée ou peu filtrée (élimination des entrées très peu fréquentes) et éventuellement de connaissances sur le contenu lexical des arguments du verbe comme nous le montrerons dans le chapitre 7. La faculté des systèmes d'acquisition automatique à s'adapter à ces besoins est l'un de leurs principaux avantages sur la constitution manuelle ou semi-manuelle de ressources.

Afin de favoriser l'utilisation de LexSchem, nous avons donc développé des outils que nous avons rendu disponibles pour la communauté. En particulier, il nous semble important de permettre aux utilisateurs potentiels d'adapter la ressource à des besoins variés.

Cette section présente ces outils développés autour d'ASSCi et de LexSchem : un programme de filtrage du lexique ; un script de transformation du format du lexique et un système de proposition de nouveaux SSC pour compléter le lexique.

6.4.1 Filtrage du lexique

Le filtrage de LexSchem réalisé par le troisième module d'ASSCi a pour objectif d'optimiser l'équilibre entre la précision et le rappel du lexique. Toutefois, selon l'application visée (lexicalisation d'un analyseur syntaxique, classification verbale, fusion des données avec d'autres ressources, ...), les informations souhaitées dans le lexique peuvent varier. Par exemple, un algorithme d'acquisition de classes verbales pourra tirer profit de schémas

18. Contrairement au verbe, on dispose de peu de dictionnaires électroniques de sous-catégorisation pour les noms et adjectifs prédicatifs en français même si certaines des ressources présentées au chapitre 3 disposent de telles informations.

contenant des modifieurs et sera peu sensible au bruit alors qu'un analyseur syntaxique aura besoin de la structure argumentale *stricto sensu*.

Un script de filtrage de LexSchem inspiré du module de filtrage d'ASSCi et le lexique non filtré sont mis à disposition de la communauté sur la page web de LexSchem : `http://www-lipn.univ-paris13.fr/~messiant/lexschem/lexschem.php` [19].

Le script fonctionne exactement de la même manière que le module de filtrage présenté dans la section 5.4.3. La seule différence est que l'utilisateur doit entrer les réglages au lancement du script.

Les choix proposés à l'utilisateur sont :

- les seuils à utiliser pour le filtrage : seuil de base, seuil pour les formes intransitives, seuil pour les formes pronominales ;
- la réduction ou non des SSC.

Après l'utilisation de ce programme, l'utilisateur peut disposer d'une ressource plus adaptée à ses besoins que le lexique présenté à la section 6.1.

6.4.2 Formats disponibles pour le lexique

Afin de rendre LexSchem utilisable dans différents contextes, celui-ci est disponible en différents formats. Le format d'origine des ressources acquises par le système ASSCi est un format « brut » : les entrées sont présentées par ligne et les champs sont séparés par des tabulations. Ce format convient bien à un traitement automatique mais il n'est pas standard.

Pour faciliter l'utilisation de nos ressources, nous avons développé et mis à disposition de la communauté un script de transformation du format de sortie d'ASSCi [20].

Le programme permet de choisir de rendre LexSchem plus lisible (format « normal »), de le transformer dans un format XML ou encore dans le format utilisé pour les versions antérieures du lexique (ce qui permet aux utilisateurs de LexSchem qui auraient utilisé ces versions d'utiliser la dernière version sans problème). Un aperçu de ces formats est donné en annexe D.

Le script permet également de faire disparaître les listes de séquences Syntex et des têtes argumentales du fichier de résultat [21] (ces informations rendent le fichier moins lisible et sont parfois inutiles).

Le format XML permet d'adapter aisément les lexiques acquis par ASSCi dans des formats standard comme le format EASy [22] ou encore le format LMF [23].

19. La version du lexique non filtré disponible ne contient pas les listes des séquences Syntex et des têtes argumentales correspondant à l'entrée (la taille du fichier contenant ces informations est trop importante pour qu'il soit facilement téléchargeable).

20. Le script de transformation et des versions de LexSchem aux différents formats sont téléchargeables sur la page : `http://www-lipn.univ-paris13.fr/~messiant/lexschem.html`.

21. Le suffixe `-noseqs` est ajouté à la fin du nom des fichiers ne contenant par les séquences Syntex et le suffixe `-noargs` à la fin du nom des fichiers ne contenant pas les têtes argumentales.

22. Un format proche d'EASy a d'ailleurs été utilisé pour comparer LexSchem à TreeLex.

23. Lexical Markup Framework : `http://www.lexicalmarkupframework.org/`

6.4.3 Proposition de nouveaux schémas

Comme nous l'avons vu dans la section 6.2.2, l'observation de notre lexique a révélé qu'un grand nombre de SSC corrects mais absents de LEXSCHEM s'avéraient être des formes réduites de schémas présents dans LEXSCHEM. Il s'agit souvent de compléments qui peuvent s'effacer en surface tout en restant présents dans la structure argumentale profonde. Par exemple, pour le verbe *alerter*, LEXSCHEM répertorie le SSC `[SUJ:SN,OBJ:SN,P-OBJ:SP<sur+SN>]` mais pas `[SUJ:SN,P-OBJ:SP<sur+SN>]`. Or, cette construction peut apparaître dans certains contextes : « *Le panneau de forme triangulaire alerte sur le danger.* ». L'objet est ici sous-entendu mais il est intéressant de disposer de ce schéma pour la lexicalisation d'un analyseur syntaxique par exemple.

Nous avons donc développé un script qui permet de compléter un lexique existant en inférant automatiquement des SSC. Ces schémas sont produits en réduisant les SSC présents dans le lexique d'origine (c'est-à-dire en supprimant un de leurs arguments). Par exemple, si le verbe *dépenser* accepte le schéma `[SUJ:SN,OBJ:SN,P-OBJ:SP<pour+SN>]` mais pas le schéma `[SUJ:SN,P-OBJ:SP<pour+SN>]`, on peut supposer qu'un effacement de l'objet est possible et la construction `[SUJ:SN,P-OBJ:SP<pour+SN>]` peut être proposée au validateur qui détermine sa pertinence. Si le SSC est validé, il est ajouté au lexique.

Pour chaque verbe du lexique (ou chacun des verbes choisis au préalable par l'utilisateur), le script parcourt ses SSC et tente de les réduire. Pour ce faire, il teste les schémas produits en retirant l'un des arguments non sujet du SSC. Si ce SSC n'est pas déjà présent dans LEXSCHEM, il est proposé à un validateur humain qui choisit de l'ajouter ou non au lexique.

Par exemple, un nouveau schéma est proposé au validateur pour le verbe *dépenser* :

> *Ajouter* `[SUJ:SN,P-OBJ:SP<pour+SN>]` *au lexique pour le verbe "dépenser" ? (ce SSC a été inféré à partir du schéma* `[SUJ:SN,OBJ:SN,P-OBJ:SP<pour+SN>]` *qui est présent dans le lexique)*

Si le validateur répond « `oui` » (en s'appuyant par exemple sur la phrase *il a dépensé pour 300 euros (en une soirée)*), le SSC est ajouté au lexique :

> *L'entrée dépenser /* `[SUJ:SN,P-OBJ:SP<pour+SN>]` *a été ajoutée au lexique !*

A l'issue du processus, on obtient une nouvelle version du lexique comprenant les ajouts validés. Néanmoins, le lexique produit est « appauvri » dans le sens où les informations relatives aux entrées (nombre d'occurrences, fréquence relative, têtes des arguments, etc) n'existent pas pour les schémas ajoutés. Le nouveau lexique est donc une suite de couples verbes/schémas.

Par exemple, le verbe *encourager* est caractérisé par les entrées suivantes dans le nouveau lexique (l'étoile marque l'entrée ajoutée grâce au programme de proposition de nouveaux schémas) :

```
dépenser [SUJ:SN]
dépenser [SUJ:SN,OBJ:SN]
dépenser [SUJ:SN,OBJ:SN,P-OBJ:SP<pour+SINF>]
dépenser [SUJ:SN,OBJ:SN,P-OBJ:SP<pour+SN>]
* dépenser [SUJ:SN,P-OBJ:SP<pour+SN>]
```

Cette méthode permet donc de compléter facilement le lexique afin d'améliorer sa couverture. Le programme peut également servir de socle pour une validation du lexique ou proposer d'autres règles pour compléter LEXSCHEM (par exemple, en proposant d'ajouter les entrées qui ont été observées dans le corpus mais dont la fréquence relative est légérement inférieure au seuil).

À l'heure actuelle, nous n'avons malheureusement pas pu mesurer l'apport de cette méthode de complétion du lexique. Néanmoins, si la validation manuelle des entrées proposées est coûteuse, elle garantit une amélioration de la ressource.

6.5 CONCLUSION

Nous avons présenté l'acquisition d'un lexique de sous-catégorisation à large couverture pour les verbes français. L'évaluation de ce lexique a permis de montrer ses limites et de tenter de répondre aux erreurs les plus fréquentes. Nous avons ensuite effectué une deuxième expérience pour mesurer l'apport des méthodes d'acquisition automatique dans le cadre d'une langue « spécialisée ». Enfin, nous avons présenté des outils développés pour faciliter l'accès et l'utilisation de ce lexique par la communauté.

Une autre manière de mesurer l'intérêt des méthodes d'acquisition automatique de ressources est d'utiliser les lexiques acquis dans un cadre applicatif. Le prochain chapitre présente des expériences de classification automatique de verbes réalisées à partir des données issues d'ASSCi.

7

PRODUCTION DE CLASSES DE VERBES SUR LA BASE DE LEUR COMPORTEMENT SYNTAXIQUE

Le chapitre précédent a montré que les lexiques acquis automatiquement comportent une part non négligeable d'erreurs. Il est néanmoins possible d'utiliser les informations présentes dans ces lexiques dans des applications de traitement automatique des langues ou pour acquérir d'autres ressources. Nous nous sommes particulièrement intéressé à la production de classes de verbes. De telles classes fournissent une base solide à la constitution de ressources telles que WordNet ou VerbNet qui ont de nombreuses applications.

Nous avons réalisé deux expériences de production automatique de classes verbales à partir de LexSchem. Ces expériences ont été réalisées dans le cadre de collaborations. Nous présentons tout d'abord la problématique et l'état de l'art du domaine avant de présenter les deux expériences de production de classes verbales et leur évaluation. Le chapitre se termine par une discussion sur les résultats obtenus.

7.1 Problématique

Les classes lexicales visent à regrouper des ensembles de mots partageant des comportements syntaxiques et/ou sémantiques similaires. Nous nous intéressons en particulier aux classes de verbes. Par exemple, on peut considérer la classe de verbes suivante :

```
[casser briser fracasser]
```

Ces verbes, proches d'un point de vue sémantique, ont également une proximité syntaxique. Ces verbes acceptent les mêmes SSC et ont des alternances communes. Dans l'exemple bien connu suivant, les schémas prédicatifs ne suffisent pas pour classifier correctement les verbes *hit* et *break* (Fillmore, 1967) :

(1a) *The boy broke the window with a ball.*
(1b) *The boy hit the window with a ball.*

Les verbes *hit* et *break* partagent la même construction : dans les deux cas, *the boy* est sujet/agent, *the window* est objet/patient et *a ball* est objet avec la préposition *with*/instrument. Or, *break* traduit un changement d'état de la fenêtre alors que ce n'est pas le cas pour *hit* (qui traduit simplement un contact sans changement d'état) et ces verbes appartiennent donc plutôt à des classes distinctes.

Ce qui permet de distinguer les verbes de changement d'état (comme break) des verbes de contact (comme *hit*), c'est le fait que les premiers acceptent l'alternance *causative/inchoative* alors que les seconds ne l'acceptent pas :

(2a) *The window broke.*
(2b) * *The window hit.*

Voici quelques exemples de verbes appartenant à ces deux classes :

```
[break shatter bend fold crack]
[hit slap strike bump stroke]
```

La classification verbale a donné lieu à de nombreux travaux (Levin, 1993; Grishman et al., 1994; Miller, 1995; Baker et al., 1998; Palmer et al., 2005; Kipper-Schuler, 2005; Hovy et al., 2006). Ces classifications, sémantiques et syntaxiques, ont de nombreux intérêts, tant applicatifs que théoriques.

D'un point de vue théorique, l'étude de ces classes permet de se pencher sur la question de l'interface syntaxe/sémantique. En effet, les travaux de Levin sur l'anglais (Levin, 1993) tout comme ceux de Gross sur le français (Gross, 1975) tendent à montrer que des verbes partageant des comportements syntaxiques similaires peuvent (souvent) former des classes homogènes sur le plan sémantique. Levin (1993) a constitué des classes de verbes sur la base des alternances qu'ils partagent : si des verbes n'ont pas les mêmes alternances, ils n'appartiennent pas à la même classe. Ensuite, Levin a tenté d'interpréter les classes obtenues en cherchant quels en sont les éléments sémantiques discriminants (par exemple, dans le cas des verbes *break* et *hit*, il s'agit du changement d'état du patient). Toutefois, ce lien entre syntaxe et sémantique n'est pas parfait. C'est pourquoi l'exploration de classes

acquises sur la base de comportements syntaxiques est intéressant pour tenter de mieux délimiter cette interface [1]

Au niveau applicatif, les classes verbales peuvent notamment être utilisées par des systèmes question/réponse, de traduction automatique, d'étiquetage en rôles sémantiques ou d'extraction d'information (Swier et Stevenson, 2004; Dang, 2004; Shi et Mihalcea, 2005; Abend et al., 2008; Zapirain et al., 2008). Nous avons également vu que des classes de ce type sont utiles pour améliorer les performances des systèmes d'acquisition de SSC en permettant de compenser un manque d'information. En effet, les verbes d'une même classe partageant d'après Levin les mêmes alternances, on peut attribuer à un verbe rare l'ensemble des SSC repérés pour un verbe moins rare de la même classe sémantique (*back-off estimates*, pour plus de détail voir la section 4.4 et l'article de Korhonen (2000)). Leur exploitation est toutefois limitée à l'heure actuelle en raison de leur faible disponibilité pour la plupart des langues.

L'intérêt de ces classes est donc manifeste pour les recherches théoriques comme pour les recherches appliquées. Il est aussi possible d'inférer de ce type de recherche une ressource « à la VerbNet », où les verbes sont assemblés en classes syntaxico-sémantiques rangées hiérarchiquement. Il a été montré qu'une approche automatique du type de celles que nous présentons dans ce chapitre constitue une base intéressante pour produire des classes pertinentes et peut notamment enrichir un travail manuel coûteux (Kipper et al., 2008).

7.2 Travaux antérieurs

Des travaux récents font état d'approches variées de classification automatique (Schulte im Walde, 2006; Joanis et al., 2008; Li et Brew, 2008; Ó Séaghdha et Copestake, 2008; Vlachos et al., 2009; Sun et Korhonen, 2009; Suzuki et Fukumoto, 2009).

L'intérêt de la classification automatique est d'une part de faciliter la tâche fastidieuse de création de ressources linguistiques par le biais d'une acquisition automatique de propriétés lexicales; d'autre part, les ressources lexicales structurées sont plus faciles à maintenir et à élargir parce qu'elles permettent de formuler des généralisations sur des classes de mots.

Les premières tentatives de classification sémantique automatique s'appuyaient sur la classification de Levin : elles visaient à reproduire automatiquement les classes anglaises ou une classification équivalente pour d'autres langues (Schulte im Walde, 2000; Korhonen et al., 2003), à identifier les alternances syntaxiques (McCarthy, 2001) ou à compléter le système de Levin par de nouvelles classes (Korhonen et Briscoe, 2004).

Nous nous intéressons ici uniquement aux méthodes non supervisées, reposant sur un espace de traits extrait d'un corpus analysé syntaxiquement

1. Il nous semble d'ailleurs que le statut de ces classes n'est pas tout à fait clair sur le plan théorique. Les approches de Levin et de Gross n'ont jamais fait l'objet d'un examen comparé approfondi. Pourtant, nous pensons que la comparaison de ces approches serait intéressante, dans la mesure où il s'agit de deux ensembles de travaux proches et importants (même s'ils reposent sur des présupposés en partie différents).

(Schulte im Walde, 2000; Schulte im Walde et Brew, 2002; Korhonen et al., 2003). Le point de départ nécessaire est donc un lexique syntaxique avec des informations concernant la fréquence relative des différents SSC par verbe, avec ou sans informations sémantiques. Ces expérimentations montrent que des distinctions syntaxiques plus détaillées ainsi que la prise en compte des modifieurs augmentent la précision de la classification (la fréquence d'apparition de certains modifieurs peut aider à identifier des classes).

Toutefois, selon Schulte im Walde (2000), l'ajout d'informations sur les restrictions de sélection conduit à un problème de manque de données (*data sparseness*) et à la baisse de performance : pour Schulte im Walde, de meilleurs résultats sont obtenus sur des espaces de traits limités à la spécification syntaxique. À l'inverse, d'autres (Alishahi et Stevenson, 2007; Li et Brew, 2008) ont essayé d'enrichir l'espace de traits de manière efficace par le biais d'informations sémantiques ou lexicales. Nous comparons dans nos expériences les résultats obtenus en utilisant diverses caractéristiques (SSC avec ou sans la différenciation des prépositions, SSC avec les têtes lexicales, etc), ce qui permettra de délimiter quelles sons les informations à fournir au système pour obtenir les meilleurs résultats.

À notre connaissance, la seule tentative pour établir une classification automatique des verbes français est décrite par Falk (2008). Cette étude s'appuie sur trois lexiques de sous-catégorisation pour comparer le comportement syntaxique des verbes. La particularité de l'approche est de partir de ressources manuelles (VOLEM, le LEXIQUE GRAMMAIRE et DICOVALENCE), ce qui permet d'utiliser des caractéristiques plus variées que les seuls SSC (caractère instrumental du sujet, alternances, etc) mais pas de prendre en considération la fréquence des SSC. Le calcul de similitude entre les verbes est fait par l'analyse formelle de concepts (Ducassé et Ferré, 2009). Il semble pourtant que les indications concernant la fréquence relative des SSC et la présence (ou non) de modifieurs soient des paramètres importants, aussi prenons-nous en compte ces éléments dans les expériences qui suivent.

7.3 DEUX EXPÉRIENCES POUR L'ACQUISITION DE CLASSES SÉMANTIQUES DE VERBES

Nous avons vu dans la section 7.1 que l'acquisition de classes verbales a des intérêts théoriques et applicatifs. Notre objectif est donc d'acquérir des classes verbales à partir de LEXSCHEM. Nous avons également vu qu'il existe plusieurs méthodes de classification et qu'un débat existe sur la quantité d'informations à fournir pour obtenir les meilleures classes possibles.

Nous avons donc réalisé en collaboration deux expériences de classification à partir de LEXSCHEM. Ces deux expériences diffèrent par la méthode utilisée et les informations exploitées, ce qui nous permet de comparer les résultats obtenus et de vérifier l'hypothèse selon laquelle ces résultats sont complémentaires.

Nous reprenons en annexe deux extraits d'articles présentant ces expériences. Ces classes sont obtenues en utilisant différentes méthodes de regroupements automatique (*clustering*). N'ayant pas développé nous-même ces algorithmes, nous avons mis en annexe F la description des expériences ainsi menées. Nous avons toutefois participé activement à la préparation des

données afin de rendre ces expériences possibles. L'expérience avec l'équipe de Cambridge a ainsi nécessité des données variées (incluant les têtes lexicales des arguments ou non, ayant été filtrées ou non, etc.), ce qui a été relativement aisé à mettre en place du fait des techniques employées pour l'acquisition. Nous avons également participé à la mise en place des expériences et à l'évaluation des résultats. Nous donnons ici un résumé de ces expériences (voir l'annexe F pour voir les extraits plus détaillés des articles publiés) et nous en discutons les résultats ensuite.

7.3.1 Expérience 1 : classification par regroupement ascendant hiérarchique

Nous avons réalisé une expérience de classification par regroupement ascendant hiérarchique en collaboration avec Kata Gábor (Messiant et al., 2010).

Pour cette expérience, nous avons utilisé la version non filtrée de LexSchem mais seules les entrées qui ont plus de 5 occurrences sont utilisées. La méthode de regroupement utilisée est la classification ascendante hiérarchique : au début du processus, chaque verbe constitue une classe et on unifie les classes de verbes les plus similaires à chaque itération. Cette méthode créée des classes disjointes (un verbe ne peut pas faire partie de deux classes), ce qui ne permet pas de tenir compte de la polysémie de certains verbes mais facilite l'interprétation des classes obtenues. L'arrêt de la classification dépend de la distance maximale entre les deux classes susceptibles d'être unifiées et de la cardinalité des classes, ce qui permet à la fois d'éviter l'absorption de nombreux verbes par quelques classes et de ne pas faire de présupposition sur le nombre de classes que l'on obtient.

Afin d'évaluer les classes obtenues, nous avons créé une classification de référence « à la main » composée de 171 verbes répartis dans 16 classes différentes[2]. Après plusieurs expérimentations, 20% des classes obtenues sont parfaitement homogènes et 43% comprennent un verbe incorrect au maximum. Ces résultats peuvent être améliorés en exploitant plus d'informations (par exemple, sur les têtes lexicales des arguments) ou en spécifiant les composants sémantiques centraux des classes (classification supervisée).

7.3.2 Expérience 2 : classification par regroupement spectral

Nous avons réalisé une autre étude de classification verbale à partir des sorties d'ASSCi (Sun et al., 2010).

La méthode de regroupement utilisée est le regroupement spectral (*spectral clustering*), suivant l'algorithme MNCut (Meila et Shi, 2001). Le nombre de classes est déterminé automatiquement en utilisant la technique proposée par Zelnik-Manor et Perona (2004). Le détail de l'implémentation est décrit dans l'article fourni en annexe F. Ce qui nous semble important ici, c'est que la variété des informations disponibles dans LexSchem a permis d'utiliser différents jeux de paramètres pour le regroupement. Un grand nombre des informations utilisées ne sont pas disponibles dans les ressources construites « à la main », ce qui montre l'intérêt des ressources acquises automatiquement pour cette tâche.

2. Cette référence a été constituée à l'aide des classes de Levin et du Lexique Grammaire.

Les jeux de paramètres utilisés sont détaillés ci-dessous. À simple titre d'illustration et pour rendre les données plus concrètes, nous donnons les informations exploitées pour le verbe *donner* (à partir de la phrase : « *Si on donnait l' avantage à un laboratoire français, ...* »[3] et des informations contenues dans LEXSCHEM).

F1 : les SSC (sans distinction des prépositions) et leurs fréquences relatives :
`[SUJ:SN, OBJ:SN, OBJ:SP<prep+SN>]; 0.35.`

F2 : idem que F1, avec une prise en compte de l'étiquette morpho-syntaxique du verbe (pour distinguer le temps verbal) :
`VCONJS; [SUJ:SN, OBJ:SN, OBJ:SP<prep+SN>]; 0.35`

F3 : idem que F2, mais en prenant en compte les prépositions utilisées au sein des compléments prépositionnels :
`VCONJS; [SUJ:SN, OBJ:SN, OBJ:SP<à+SN>]; 0.35.`

Les six jeux de paramètres suivants prennent en compte le contexte lexical des verbes. Les colocations sont extraites à partir des fenêtres gauche et droite du verbe (en ignorant les mots vides).

F4, F6, F8 : les contextes de 4, 6 et 8 mots. La position du mot est ignorée :
`{ avantage, à }` (4 mots).

F5, F7, F9 : idem que F4, F6 et F8 avec un enregistrement de la position relative du mot :
`gauche : ∅, droite : { avantage, à, laboratoire }` (6 mots).

Les quatre jeux de paramètres suivants utilisent les préférences lexicales (lp) des verbes dans certaines positions argumentales (lemmes têtes) :

F10 : LP(PREP) : les prépositions acceptées par le verbe et leur fréquence :
`(à : 1).`

F11 : LP(SUBJ) : le type et la fréquence des noms dans la relation *sujet* :
`(il : 0.09, on : 0.04, ...)` (« *il donne...* », « *on donne...* », etc).

F12 : LP(IOBJ) : le type et la fréquence des noms dans les relations *objet* et *objet indirect* :
`(nom : 0.03, coup : 0.03, ...)` pour la relation *objet* (« *donner un nom* », « *donner un coup* », etc)
et `(lui : 0.20, gouvernement : 0.01, ...)` pour la relation *objet indirect* (« *lui donner* », « *donner au gouvernement* », etc).

F13 : LP(ALL) : la combinaison des caractéristiques F10, F11 et F12.

Les deux derniers jeux de paramètres combinent les SSC avec les préférences lexicales et des restrictions de sélection :

F14-F16 : idem que F1-F3, avec une prise en compte des préférences lexicales : VCONJS;

```
[ SUJ:SN (il : 0.09, on : 0.04, ...),
  OBJ:SN (nom : 0.03, coup : 0.03, ...),
  OBJ:SP<à+SN (lui : 0.20, gouvernement : 0.01, ...)>
] ; 0.35.
```

3. extrait du corpus *LM10*

F17 : F3 combinée avec des restrictions de sélection sur les arguments (40 restrictions de sélection ont été acquises automatiquement par classification non supervisée sur les têtes lexicales).

Le *gold standard* utilisé pour l'évaluation est le même que celui que nous avons utilisé dans l'expérience 1 (section 7.3.1). Une première expérience a été réalisée en se limitant aux 116 verbes qui apparaissent au moins 150 fois dans le corpus en raison de la taille critique nécessaire pour la classification automatique. Les jeux de paramètres F1-F3 donnent des résultats significativement meilleurs que la *baseline*. Parmi ces jeux de paramètres, F3 donne les meilleurs résultats et sera utilisée comme base pour F14-F17. Les jeux de paramètres qui exploitent les colocations (F4-F9) et celle qui exploite les préférences lexicales sur tous les arguments (F13) obtiennent de meilleures performances que celles qui utilisent les SSC mais c'est la combinaison des SSC et des préférences lexicales (F14-F16), voire des restrictions de sélection (F17) qui semble être la plus efficace. Toutefois, les différences de performances ne sont pas significatives. Nous avons alors réalisé une autre expérience qui montre que le jeu de paramètres le plus sophistiqué (F17) est de plus en plus efficace avec l'augmentation du nombre d'exemples considérés alors que les autres jeux de paramètres ne tirent pas autant profit de ces données.

Une évaluation qualitative des classes obtenues a aussi été réalisée et permet d'identifier les principales sources d'erreurs : faible fréquence de certaines constructions ou de certains verbes dans le corpus, polysémie, ... Elle a également permis de montrer que la classification sépare des antonymes qui font partie de la même classe dans la référence (par exemple, *affaiblir* et *renforcer*) ou regroupe des classes proches syntaxiquement et sémantiquement mais qui sont distinctes dans la référence. Dans ce cas, on ne peut pas vraiment parler d'erreurs dans le sens où l'information acquise est valide et utile.

Cette étude a révélé qu'un système optimisé pour la classification verbale en anglais, tant du point de vue méthodologique que du point de vue des caractéristiques utilisées, obtenait de bons résultats pour le français. Elle a aussi mis en évidence l'importance de la taille des données utilisées pour ce type de travaux.

7.4 Discussion

Les résultats de ces deux expériences montrent que la classification automatique de verbes sur la base de leurs comportements syntaxiques (SSC) donne des classes cohérentes au niveau sémantique quand on les compare à un *gold standard*. De plus, ces études « valident » l'utilisation de SSC acquis automatiquement dans un cadre applicatif malgré le caractère imparfait de la ressource. La disponibilité d'informations issues du corpus (fréquences des entrées, têtes lexicales) est même un atout pour des travaux de ce type.

Les deux expériences ont également permis de vérifier l'hypothèse qu'une technique de classification utilisée sur une langue donnée peut être utilisée sur une autre langue à condition de prendre en compte les spécificités de la langue visée (par exemple, pour le français, les verbes pronominaux et la diversité des compléments prépositionnels).

Nous avons réalisé deux expériences distinctes afin de pouvoir comparer les résultats obtenus en fonction de la méthode utilisée et des informations exploitées. L'hypothèse de la complémentarité des résultats obtenus se vérifie en partie. Les deux techniques de classification utilisées sont distinctes et produisent des résultats différents. Le choix de la technique de classification a des implications directes sur les classes obtenues : la première expérience (Messiant et al., 2010) fournit des classes de taille fixe (c'est-à-dire que toutes les classes obtenues regroupent le même nombre de verbes) tandis que la deuxième expérience (Sun et al., 2010) permet d'obtenir des classes de taille variable. Si les classes de taille variable rendent mieux compte de la réalité des phénomènes linguistiques (il n'y a aucune raison pour que toutes les classes de verbes aient la même taille), elles sont difficiles à valider étant donné les frontières floues qui les délimitent. Les classes à taille fixe forment des « noyaux de classes » cohérents qui constituent une base intéressante pour un travail manuel qui permettra de regrouper ou de séparer ces classes afin d'obtenir une ressource de qualité optimale. Une perspective de ces expériences est donc de « mixer » les deux approches : en partant des noyaux de classes à taille fixe, on peut guider l'ajout de verbes *via* l'autre méthode.

L'autre principale différence réside dans l'utilisation d'une grande variété de caractéristiques dans la classification par regroupement spectral (prépositions, types et fréquences des têtes lexicales, etc). Nous avons montré que l'utilisation d'un grand nombre de caractéristiques permet une amélioration des résultats à condition de disposer de données de taille suffisante (plusieurs milliers d'occurrences par verbe).

Si l'intérêt applicatif des classes sémantiques est indéniable, ces expériences induisent également une réflexion théorique sur la nature des classes obtenues : que représentent ces classes ? que disent-elles sur la langue ?

7.5 Conclusion

Nous avons réalisé deux expériences de classification verbale à partir des sorties d'ASSCi. Ces travaux ont permis de montrer que les informations acquises par notre système permettent d'obtenir des classes verbales cohérentes sémantiquement. Ces résultats soulignent donc l'intérêt de l'acquisition automatique d'informations lexicales à partir de corpus. En effet, les expériences présentées ici n'auraient pas pu être réalisées avec des lexiques construits manuellement, même « idéaux », c'est-à-dire idéalement complets et robustes, ce qui n'est pas le cas des dictionnaires de l'état de l'art. Dans le cadre d'expériences de ce type, la plus value apportée par les techniques d'acquisition automatique, en particulier celles des informations de fréquence des SSC et de lemmes têtes, est remarquable.

CONCLUSION

8

8.1 RAPPEL DES ENJEUX

Dans l'introduction de cette thèse, nous avions identifié trois enjeux importants concernant l'acquisition automatique d'informations lexicales à partir de corpus :

1. L'étude des techniques d'acquisition automatique à large couverture pour le français, peu étudiées jusqu'alors ;
2. La validation de ces méthodes et de leur apport par rapport aux ressources construites suite à un travail manuel (tant sur le domaine général que sur un domaine de spécialité) ;
3. L'exploration de l'interface syntaxe - sémantique, notamment à travers l'acquisition de classes verbales à partir du lexique acquis.

Nous nous sommes intéressés en particulier aux schémas de sous-catégorisation de verbes qui constituent un bon point de départ pour l'acquisition d'autres informations lexicales et dont les applications sont variées : analyse syntaxique, classification verbale, etc.

8.2 CONTRIBUTION

Pour répondre à ces enjeux, nous avons montré dans cette thèse l'intérêt de l'acquisition automatique d'informations lexicales à partir de corpus bruts. Nous avons en particulier montré qu'il était possible d'acquérir automatiquement des informations de sous-catégorisation verbale pour le français et exploré des pistes d'utilisation des schémas de sous-catégorisation obtenus. Pour cela, nous avons :

- développé un système d'acquisition automatique de SSC à partir de corpus brut pour le français ;
- utilisé ce système pour acquérir un lexique de sous-catégorisation à large couverture pour le français appelé LEXSCHEM ;
- évalué les informations acquises, tant quantitativement que qualitativement ;
- montré que l'utilisation d'ASSCi sur un corpus « marqué » par un domaine (législatif européen) donnait des résultats sensiblement différents ;
- proposé des outils pour faciliter l'utilisation de telles ressources ;
- réalisé deux expériences de classification automatique de verbes à partir de nos données et montré que les classes obtenues avaient une portée sémantique.

Depuis le début des années 90, de nombreux travaux d'acquisition automatique de SSC ont été réalisés, en particulier pour l'anglais. Toutefois, en raison de la faible disponibilité de corpus et d'analyseurs syntaxiques robustes, il existait peu d'études de ce type sur la langue française au moment où cette thèse a débuté. Le projet ANR Passage, qui réunit plusieurs laboratoires et qui a été lancé au début 2007 a abouti à une étude du même type que la nôtre (Gardent et Lorenzo, 2010), ce qui montre l'actualité et l'importance de cet enjeu.

Le principal apport de cette thèse est la proposition d'une méthode d'acquisition automatique de SSC verbaux à partir de corpus bruts en français. Cette méthode est originale par plusieurs aspects : 1) elle concerne le français ; 2) elle n'utilise pas de liste de SSC *a priori* ; 3) elle prend en compte le problème de la distinction argument/modifieur et propose des pistes pour y répondre.

Un système d'acquisition de SSC nommé ASSCi a été développé à partir de cette méthode. S'il s'appuie sur l'étiqueteur morphosyntaxique TreeTagger et sur l'analyseur syntaxique Syntex, la méthode n'est pas dépendante de ces outils et peut être implémentée pour d'autres analyseurs.

La distinction automatique des arguments et des modifieurs est un problème difficile. Il serait intéressant d'approfondir cette question par exemple en utilisant la méthode proposée par (Fabre et Bourigault, 2008). Cette méthode repose sur la productivité des couples verbe/préposition, c'est-à-dire la diversité des contextes nominaux dans lesquels ce couple apparaît : plus un couple verbe/préposition apparaît avec des noms variés, plus le complément prépositionnel a de chances d'être un argument.

Notre travail constitue toutefois l'un des premiers travaux concernant l'acquisition automatique de SSC à partir de corpus brut pour le français réalisé en largeur et ayant fait l'objet d'une évaluation rigoureuse.

Ce système a permis d'acquérir un lexique de sous-catégorisation à large couverture pour le français appelé LexSchem.

LexSchem contient 7 239 entrées (couples verbe-SSC). Ces entrées concernent 3 123 lemmes verbaux et 88 SSC différents.

De plus, cette ressource est dotée d'une interface de consultation et d'outils de paramétrage qui permettent d'adapter le lexique en fonction des besoins applicatifs.

Comparée à d'autres ressources pour le français, la couverture de LexSchem (en nombre de verbes) est dans la fourchette basse. La ressource concerne toutefois les verbes les plus communs du français. En outre, il est possible d'améliorer sa couverture en augmentant les données d'acquisition (années 2001 à 2010 du journal *Le Monde*, corpus web, autres corpus) ou en compensant les informations manquantes à l'aide de classes verbales par la technique proposée par Korhonen (2000) (*back-off estimates*).

Nous avons alors évalué cette ressource, tant quantitativement que qualitativement. L'évaluation quantitative a consisté à montrer l'écart entre LexSchem et des ressources construites manuellement. En particulier, nous avons montré que LexSchem apporte une certaine nouveauté par rapport à TreeLex et DicoValence.

La comparaison des informations acquises avec les informations contenues dans les lexiques existants a aussi mis en exergue les avantages et les

limites de chacune des approches et de montrer ce que les approches automatiques apportent à l'état de l'art. Les limites de l'acquisition automatique portent principalement sur la quantité d'informations erronées acquises (en raison d'erreurs d'analyse ou de filtrage et de la difficulté à traiter la distinction argument/modifieur. Parmi les apports des techniques d'acquisition automatique, il nous semble que leur adaptabilité est le plus notable et prometteur.

En effet, nous avons montré que les SSC acquis varient en fonction du corpus utilisé, particulièrement pour les corpus de langue dite « spécialisée ». L'utilisation d'ASSCi sur le corpus *EuroParl* a permis d'acquérir un lexique de sous-catégorisation spécialisé pour le domaine législatif européen (LexSchem-EuroParl). Ce lexique est composé de 2 703 entrées (1 171 lemmes verbaux et 74 SSC différents).

Cette étude préliminaire a permis de prouver que lors d'études concernant un domaine particulier, l'utilisation de ressources acquises automatiquement peut se révéler plus intéressante que celle de dictionnaires construits manuellement.

Dans cette étude, nous n'avons comparé LexSchem-EuroParl qu'avec LexSchem. La comparaison des entrées présentes dans LexSchem-EuroParl avec les entrées présentes dans les autres lexiques de langue générale (DicoValence ou le Lexique Grammaire par exemple) serait profitable car elle permettrait de mieux mettre en évidence l'intérêt des méthodes automatiques dans le cas des « langues de spécialité ».

Une approche combinée (acquisition automatique dans un premier temps puis validation manuelle) semble idéale pour obtenir une ressource de bonne qualité.

Enfin, cette thèse s'attache à prouver que les ressources acquises, bien qu'imparfaites, peuvent être utilisées dans un cadre applicatif ou pour acquérir d'autres ressources. Dans certains contextes, il s'avère même qu'elles ont une « plus value » sur les dictionnaires existants car elles sont dotées d'informations (notamment sur la fréquence des entrées ou les lemmes têtes rencontrés en corpus) dont ne disposent pas ces lexiques. Les deux expériences de classification automatique de verbes à partir de LexSchem présentées au chapitre 7 ont permis de vérifier cette hypothèse. Elles ont aussi montré que les classes obtenues avaient une portée sémantique et procurent une base pour l'étude de l'interface syntaxe-sémantique. Néanmoins, par manque de temps, une étude approfondie des liens entre syntaxe et sémantique n'a pas pu être menée.

L'utilisation de LexSchem dans une application d'extraction d'information ou un analyseur syntaxique reste à faire. L'évaluation de l'apport de la ressource à l'application n'est pas aisée et une telle étude dépasse le cadre de cette thèse.

8.3 Perspectives

La thèse répond donc bien aux enjeux importants que nous avons identifiés. Néanmoins, les contraintes de temps et de moyens ne nous ont pas per-

mis d'explorer en profondeur certains aspects. Nous proposons ici quelques unes de ces pistes.

Tout d'abord, il est possible d'améliorer la méthode d'acquisition. Les améliorations les plus naturelles concernent la technique de filtrage employée. À l'instar de ce qui a été fait pour l'anglais (Korhonen et al., 2000), une comparaison de différentes techniques pourrait être réalisée. Le test binomial a déjà été utilisé pour le français (Chesley et Salmon-Alt, 2006) et a donné des résultats proches des nôtres. Il serait également intéressant d'essayer d'utiliser des techniques de filtrage tout à fait inédites pour cette tâche. Maintenant que l'on dispose de classes de verbes, une expérience de compensation de manques d'informations en corpus pour certains verbes, analogue à celle menée à Cambridge (*back-off estimates*) (Korhonen, 2000), pourrait être réalisée afin d'augmenter la couverture de la ressource. Une telle étude nécessite d'utiliser des classes verbales acquises à partir d'un lexique très peu filtré. Outre les progrès résultant de ces modifications de la méthode, l'amélioration des performances du système d'acquisition est également envisageable. Cela peut notamment passer par l'utilisation d'un étiqueteur morpho-syntaxique plus efficace ou par un réentrainement de TreeTagger.

Le programme de proposition de nouveaux SSC pourrait proposer des schémas issus d'autres ressources ou qui ont été filtrés alors qu'ils sont présents assez fréquemment dans le corpus. L'inconvénient principal de cet ajout est qu'il augmente significativement la part « manuelle » dans la constitution du lexique. Cependant, il nous semble que la complémentarité des approches est nécessaire pour obtenir une ressource de très bonne qualité en un temps acceptable.

Une autre extension possible de la thèse est la poursuite et l'approfondissement des expériences qui concernent l'adaptabilité de la méthode d'acquisition. Nous pensons notamment à l'acquisition de SSC pour d'autres domaines de spécialité (par exemple, le domaine médical pour lequel les besoins sont nombreux) ou à l'acquisition automatique de SSC de noms prédicatifs et d'adjectifs prédicatifs. Nous avons montré dans cette thèse que de telles adaptations étaient possibles et peu coûteuses, à condition de disposer de corpus de taille suffisante.

À plus long terme, une piste intéressante serait d'enrichir LexSchem pour en faire un lexique de schémas prédicatifs, c'est-à-dire de schémas comprenant non seulement les informations de sous-catégorisation mais aussi les rôles thématiques et les restrictions de sélection sur les arguments. L'étiquetage de rôles thématiques a fait l'objet de travaux récents (Gildea et Jurafsky, 2002). Il faut tout d'abord identifier le nombre et la granularité des rôles. Cette tâche nécessite des informations sur la sémantique profonde des verbes. Un important travail manuel est donc requis avant d'automatiser le processus. La grande majorité des études concernant l'ajout de restrictions de sélection sur les arguments utilisent des ressources du type de WordNet (Fellbaum, 1998) afin d'exploiter l'hyperonymie à partir des têtes lexicales des arguments (Resnik, 1997; Li et Abe, 1998; Abney et Light, 1999; Clark et Weir, 2002). La présence dans les ressources acquises par ASSCi des lemmes têtes des arguments est donc un avantage considérable pour acquérir automatiquement ces informations. Toutefois, il n'existe pas de ressource aussi

aboutie que WordNet disponible pour le français malgré des travaux récents allant dans ce sens comme EuroWordNet ou WOLF (Sagot et Fiser, 2008).

Enfin, l'exploration des applications directes des schémas obtenus est un enjeu majeur pour le futur. Ces applications vont de la lexicalisation d'un analyseur syntaxique avec LexSchem à l'utilisation des SSC pour l'extraction d'information. L'étude de Carroll et al. (1998) a montré que la lexicalisation d'un analyseur syntaxique avec des SSC acquis automatiquement améliore les performances de l'analyseur. L'intégration de structure argumentale dans les systèmes d'extraction d'information a aussi fait preuve de bons résultats (Surdeanu et al., 2003). Cependant, nous avons vu que l'évaluation de l'apport de la ressource est délicate. Par exemple, il est difficile de distinguer ce qui provient du système d'extraction de ce qui est issu du lexique. Ceci exige donc des protocoles expérimentaux très précis et délicats à mettre en place.

À l'issue de cette étude, nous espérons avoir montré l'intérêt et l'importance des méthodes d'acquisition automatique de connaissances lexicales pour le TAL. Cette thèse a exploré différentes pistes pour le français mais de nombreux travaux seront encore nécessaires pour améliorer les techniques existantes. Leur utilisabilité et leur complémentarité avec les méthodes manuelles sont des aspects cruciaux à approfondir à l'avenir.

Liste des étiquettes de TreeTagger pour le français

A

ABR	Abréviation
ADJ	Adjectif
ADV	Adverbe
DET :ART	Article
DET :POS	Pronom Possessif (ma, ta, ...)
INT	Interjection
KON	Conjonction
NAM	Nom Propre
NOM	Nom
NUM	Numéral
PRO	Pronom
PRO :DEM	Pronom Démonstratif
PRO :IND	Pronom Indéfini
PRO :PER	Pronom Personnel
PRO :POS	Pronom Possessif (mien, tien, ...)
PRO :REL	Pronom Relatif
PRP	Préposition
PRP :det	Préposition + Article (au, du, aux, des)
PUN	Ponctuation
PUN :cit	Ponctuation de citation
SENT	Balise de phrase
SYM	Symbole
VER :cond	Verbe au conditionnel
VER :futu	Verbe au futur
VER :impe	Verbe à l'impératif
VER :impf	Verbe à l'imparfait
VER :infi	Verbe à infinitif
VER :pper	Verbe au participe passé
VER :ppre	Verbe au participe présent
VER :pres	Verbe au présent
VER :simp	Verbe au passé simple
VER :subi	Verbe à l'imparfait du subjonctif
VER :subp	Verbe au présent du subjonctif

Exemple d'analyse syntaxique au format de Syntex

B

L'exemple présenté ici correspond à la séquence « *Les fameuses « erreurs » de Sartre, il les lui reproche autant que ses adversaires, mais il les lui reproche au nom du Sartre qu'il aime, et qui lui ressemble, l' homme seul, le Roquentin.* » (figures 5.4 et 5.5, section 5.2.2). Les éléments sont ici présentés par ligne (dans l'analyse renvoyée par Syntex, les éléments sont séparés par des tabulations). Cet extrait est donné à titre indicatif, afin d'illustrer le format utilisé par Syntex. Des explications sur les informations fournies sont données dans le corps de la thèse, au chapitre 5 (section 5.2.2, page 62).

```
Det??|le|Les|1|DET;4|
AdjFP|fameux|fameuses|2|ADJ;4|
Typo|«|«|3||
Nom?P|erreur|erreurs|4||DET;1,ADJ;2,PREP;6
Typo|»|»|5||
Prep|de|de|6|PREP;4|NOMPREP;7
NomPrXXInc|Sartre|Sartre|7|NOMPREP;6|
Typo|,|,|8||
Pro|il|il|9|SUJ;12|
Pro|le|les|10|OBJ;12|
Pro|lui|lui|11|PREP;12|
VCONJS|reprocher|reproche|12||SUJ;9,OBJ;10,PREP;11
CSub|autant que|autant que|13||
DetMP|son|ses|14|DET;15|
Nom?P|adversaire|adversaires|15||DET;14
Typo|,|,|16||
CCoord|mais|mais|17||
Pro|il|il|18|SUJ;21|
Pro|le|les|19|OBJ;21|
Pro|lui|lui|20|PREP;21|
VCONJS|reprocher|reproche|21||SUJ;18,OBJ;19,PREP;20,PREP;22
Prep|au nom de|au nom du|22|PREP;21|NOMPREP;23
NomPrXXInc|Sartre|Sartre|23|NOMPREP;22|
CSub|que|qu'|24||COMP;26
Pro|il|il|25|SUJ;26|
VCONJS|aimer|aime|26|COMP;24|SUJ;25,OBJ;34
Typo|,|,|27||
CCoord|et|et|28||
```

```
ProRel|qui|qui|29|SUJ;31|
Pro|lui|lui|30|PREP;31|
VCONJS|ressembler|ressemble|31||SUJ;29,PREP;30
Typo|,|,|32||
Det??|le|l'|33|DET;34|
Nom?S|homme|homme|34|OBJ;26|DET;33,ADJ;35
Adj??|seul|seul|35|ADJ;34|
Typo|,|,|36||Det??|le|le|37|DET;38|
NomMS|roquentin|Roquentin|38||DET;37
Typo|.|.|39||
```

Liste des prépositions issues de PrepLex

C

Nous donnons ici la liste des prépositions argumentales issues de PrepLex, c'est-à-dire l'ensemble des prépositions susceptibles d'introduire un argument autour du verbe, puis la liste des prépositions non-argumentales (voir `http://loriatal.loria.fr/Resources/PrepLex.txt` pour la liste complète des locutions non-argumentales).

Prépositions argumentales :

à	par	vis-à-vis de
de	sur	face à
après	parmi	le long de
avec	pour	du côté de
chez	sans	au-dessus de
comme	selon	au-devant de
contre	sous	autour de
dans	suivant	envers
depuis	vers	en faveur de
derrière	à travers	auprès de
devant	à partir de	jusqu'à
en	au sujet de	d'avec
entre	au travers de	

Prépositions non-argumentales :

avant	circa	concernant
confer	dès	dixit
durant	excepté	ès
hormis	hors	jusque
malgré	moyennant	nonobstant
outre	passé	pendant
sauf	sitôt	touchant
versus	via	voici
voilà	vu	

D

ENTRÉES DE LEXSCHEM POUR LE VERBE *accompagner* DANS LES DIFFÉRENTS FORMATS DISPONIBLES

Les deux entrées du verbe *accompagner* sont présentées dans les formats disponibles pour LEXSCHEM (format « standard », et format XML). Pour une meilleure lisibilité, les listes des identifiants des séquences du corpus et les lemmes têtes des arguments ont été tronquées (les versions complètes du lexique sont disponibles à l'adresse suivante : `http://www-lipn.univ-paris13.fr/~messiant/lexschem.html`).

FORMAT « STANDARD »

```
*** Entrée 04306 (6135 occurences, fréquence relative: 75.4)
Verbe : accompagner (8142 occurences, 2 ssc)
SSC : [SUJ:SN,OBJ:SN]
Séquences Syntex : 0100.anasynt!d684217p3_5!38,
  0100.anasynt!d687293p5_4!19, 0100.anasynt!d683973p3_3!39,
  ...
Lemmes têtes de l'argument 0 :  #UNKNOWN (527:0.08),
  il (477:0.08), qui (384:0.06), ...
Lemmes têtes de l'argument 1 :  le (1699:0.27),
  mouvement (68:0.01), exposition (66:0.01), ...

*** Entrée 00239  (2007 occurences, fréquence relative: 24.6)
Verbe : accompagner (8142 occurences, 2 ssc)
SSC : [SUJ:SN,REF:refl,DE-OBJ:SP<de+SN>]
Séquences Syntex : 0100.anasynt!d686200p4_17!48,
  0100.anasynt!d686690p5_6!16, 0100.anasynt!d686339p4_30!3,
  ...
Lemmes têtes de l'argument 0 :  #UNKNOWN (215:0.11),
  elle (212:0.10), il (108:0.05), ...
Lemmes têtes de l'argument 1 :  se (2007:1.00)
Lemmes têtes de l'argument 2 :  baisse (38:0.02),
  augmentation (31:0.02), réduction (28:0.01), ...
```

FORMAT XML

```
<lexschem>
   ...
   <entree id="04306" >
      <verbe nb_occ="8142" nb_ssc="2" >
         ACCOMPAGNER+accompagner
      </verbe>
      <schema>
         [<SUJ:SN,OBJ:SN>]
      </schema>
      <nb_occ>
         6135
      </nb_occ>
      <rel_freq>
         0.754
      </rel_freq>
      <sequences>
         0100.anasynt!d684217p3_5!38,
         0100.anasynt!d687293p5_4!19,
         0100.anasynt!d683973p3_3!39, ...
      </sequences>
      <nb_args>
         2
      </nb_args>
      <argument0>
         #UNKNOWN (527:0.08), il (477:0.08), qui (384:0.06),
         ...
      </argument0>
      <argument1>
         le (1699:0.27), mouvement (68:0.01), exposition
         (66:0.01), ...
      </argument1>
   </entree>
   <entree id="00239" >
      <verbe nb_occ="8142" nb_ssc="2" >
         ACCOMPAGNER+accompagner
      </verbe>
      <schema>
         [<SUJ:SN,REF:refl,DE-OBJ:SP<de+SN>>]
      </schema>
      <nb_occ>
         2007
      </nb_occ>
      <rel_freq>
         0.246
      </rel_freq>
      <sequences>
         0100.anasynt!d686200p4_17!48,
         0100.anasynt!d686690p5_6!16,
         0100.anasynt!d686339p4_30!3, ...
      </sequences>
      <nb_args>
         3
      </nb_args>
      <argument0>
         #UNKNOWN (215:0.11), elle (212:0.10), il (108:0.05),
```

```
            ...
         </argument0>
         <argument1>
            se (2007:1.00)
         </argument1>
         <argument2>
            baisse (38:0.02), augmentation (31:0.02),
            réduction (28:0.01), ...
         </argument2>
      </entree>
      ...
</lexschem>
```

Liste des SSC de LexSchem

E

Schéma de sous-catégorisation	Nombre d'occurrences dans le corpus *LM10*
[SUJ:SN,OBJ:SN]	2597802
[SUJ:SN]	1621286
[SUJ:SN,OBJ:SINF]	1008112
[SUJ:SN,OBJ:PropSub]	444432
[SUJ:SN,REF:refl]	405430
[SUJ:SN,DE-OBJ:SP<de+SINF>]	229668
[SUJ:SN,A-OBJ:SP<à+SN>]	226368
[SUJ:SN,OBJ:SN,A-OBJ:SP<à+SN>]	161980
[SUJ:SN,A-OBJ:SP<à+SINF>]	136260
[SUJ:SN,DE-OBJ:SP<de+SN>]	131627
[SUJ:SN,ATTS:SA]	76086
[SUJ:SN,REF:refl,DE-OBJ:SP<de+SN>]	55921
[SUJ:SN,ATTS:SN]	47221
[SUJ:SN,REF:refl,A-OBJ:SP<à+SN>]	41675
[SUJ:SN,P-OBJ:SP<sur+SN>]	38800
[SUJ:SN,P-OBJ:SP<dans+SN>]	37829
[SUJ:SN,A-OBJ:SP<à+SN>,DE-OBJ:SP<de+SINF>]	29399
[SUJ:SN,ATTS:SINF]	28900
[SUJ:SN,REF:refl,DE-OBJ:SP<de+SINF>]	25190
[SUJ:SN,OBJ:SN,A-OBJ:SP<à+SINF>]	22503
[SUJ:SN,REF:refl,A-OBJ:SP<à+SINF>]	19560
[SUJ:SN,OBJ:SINF,OBJ:SN]	15152
[SUJ:SN,OBJ:SN,P-OBJ:SP<dans+SN>]	14888
[SUJ:SN,P-OBJ:SP<avec+SN>]	12416
[SUJ:SN,OBJ:SN,DE-OBJ:SP<de+SINF>]	11577
[SUJ:SN,REF:refl,P-OBJ:SP<sur+SN>]	11523
[SUJ:SN,OBJ:SN,ATTO:SA]	10111
[SUJ:SN,P-OBJ:SP<par+SINF>]	9631
[SUJ:SN,OBJ:SN,DE-OBJ:SP<de+SN>]	9564
[SUJ:SN,OBJ:SN,P-OBJ:SP<sur+SN>]	9525
[SUJ:SN,REF:refl,P-OBJ:SP<dans+SN>]	8638
[SUJ:SN,OBJ:SN,P-OBJ:SP<en+SN>]	8498
[SUJ:SN,P-OBJ:SP<pour+SN>]	7916
[SUJ:SN,P-OBJ:SP<en+SN>]	6828
[SUJ:SN,OBJ:SN,P-OBJ:SP<avec+SN>]	6819
[SUJ:SN,OBJ:SN,P-OBJ:SP<comme+SN>]	6053
[SUJ:SN,REF:refl,OBJ:SN]	5974
[SUJ:SN,P-OBJ:SP<par+SN>]	5621

[SUJ:SN,REF:refl,ATTO:SA]	4926
[SUJ:SN,P-OBJ:SP<comme+SN>]	4882
[SUJ:SN,OBJ:SN,P-OBJ:SP<pour+SINF>]	4864
[SUJ:SN,A-OBJ:SP<à+SN>,DE-OBJ:SP<de+SN>]	4796
[SUJ:SN,REF:refl,P-OBJ:SP<par+SN>]	4742
[SUJ:SN,REF:refl,OBJ:PropSub]	3923
[SUJ:SN,P-OBJ:SP<contre+SN>]	3242
[SUJ:SN,REF:refl,P-OBJ:SP<en+SN>]	3114
[SUJ:SN,OBJ:SN,P-OBJ:SP<pour+SN>]	2732
[SUJ:SN,DE-OBJ:SP<de+SN>,P-OBJ:SP<pour+SINF>]	1942
[SUJ:SN,P-OBJ:SP<pour+SINF>]	1787
[SUJ:SN,DE-OBJ:SP<de+SN>,P-OBJ:SP<en+SN>]	1765
[SUJ:SN,OBJ:SN,P-OBJ:SP<depuis+SN>]	1446
[SUJ:SN,P-OBJ:SP<parmi+SN>]	1396
[SUJ:SN,P-OBJ:SP<autour de+SN>]	1297
[SUJ:SN,P-OBJ:SP<entre+SN>]	1248
[SUJ:SN,OBJ:SN,P-OBJ:SP<par+SN>]	1157
[SUJ:SN,REF:refl,P-OBJ:SP<avec+SN>]	966
[SUJ:SN,P-OBJ:SP<depuis+SN>]	907
[SUJ:SN,REF:refl,P-OBJ:SP<contre+SN>]	858
[SUJ:SN,REF:refl,P-OBJ:SP<vers+SN>]	809
[SUJ:SN,P-OBJ:SP<vers+SN>]	722
[SUJ:SN,P-OBJ:SP<devant+SN>]	684
[SUJ:SN,P-OBJ:SP<en faveur de+SN>]	627
[SUJ:SN,P-OBJ:SP<chez+SN>]	593
[SUJ:SN,REF:refl,P-OBJ:SP<pour+SN>]	484
[SUJ:SN,OBJ:SN,P-OBJ:SP<vers+SN>]	399
[SUJ:SN,A-OBJ:SP<à+SN>,P-OBJ:SP<pour+SINF>]	335
[SUJ:SN,P-OBJ:SP<selon+SN>]	333
[SUJ:SN,REF:refl,P-OBJ:SP<devant+SN>]	306
[SUJ:SN,P-OBJ:SP<sous+SN>]	257
[SUJ:SN,P-OBJ:SP<après+SN>]	233
[SUJ:SN,OBJ:SN,P-OBJ:SP<contre+SN>]	211
[SUJ:SN,REF:refl,P-OBJ:SP<autour de+SN>]	201
[SUJ:SN,REF:refl,P-OBJ:SP<derrière+SN>]	180
[SUJ:SN,REF:refl,P-OBJ:SP<pour+SINF>]	152
[SUJ:SN,REF:refl,P-OBJ:SP<entre+SN>]	137
[SUJ:SN,P-OBJ:SP<pour+SINF>,P-OBJ:SP<sur+SN>]	100
[SUJ:SN,REF:refl,A-OBJ:SP<à+SN>,DE-OBJ:SP<de+SN>]	77
[SUJ:SN,OBJ:SN,P-OBJ:SP<entre+SN>]	73
[SUJ:SN,DE-OBJ:SP<de+SN>,P-OBJ:SP<entre+SN>]	72
[SUJ:SN,OBJ:SN,A-OBJ:SP<à+SN>,DE-OBJ:SP<de+SINF>]	48
[SUJ:SN,P-OBJ:SP<au-dessus de+SN>]	47
[SUJ:SN,P-OBJ:SP<du côté de+SN>]	28
[SUJ:SN,OBJ:SN,P-OBJ:SP<contre+SN>,P-OBJ:SP<en+SN>]	24
[SUJ:SN,OBJ:SN,P-OBJ:SP<devant+SN>]	16
[SUJ:SINF,OBJ:SN]	16
[SUJ:SN,P-OBJ:SP<à_travers+SN>]	14
[SUJ:SN,A-OBJ:SP<à+SN>,P-OBJ:SP<dans+SN>]	12
[SUJ:SN,OBJ:SN,P-OBJ:SP<jusqu'à+SN>]	10

ARTICLES PORTANT SUR L'ACQUISITION DE CLASSES VERBALES

F

Comme nous n'avons pas développé nous-même les algorithmes de classification, nous donnons en annexe le texte des expériences faites en commun avec Kata Gábor d'une part et l'équipe de Cambridge d'autre part. Une présentation de ces expériences et une discussion des résultats figure dans le chapitre 7 de la thèse.

F.1 EXTRAIT DE L'ARTICLE PUBLIÉ DANS LA REVUE TAL EN 2010

Cette section est un extrait de l'article :

Cédric Messiant, Kata Gábor et Thierry Poibeau. Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français. *Traitement Automatique des Langues*, 2010.

F.1.1 Méthode

Dans le cadre de cette expérimentation, nous nous sommes intéressés à la classification automatique de verbes français en classes lexico-sémantiques. Pour ce faire, nous partons de l'hypothèse qu'il est possible de produire ce type d'information par des méthodes statistiques appliquées à des données syntaxiques. L'hypothèse est fondée sur l'observation qu'il existe une corrélation entre les propriétés sémantiques des verbes et leurs contextes syntaxiques. Levin (1993) a donné une description systématisée du phénomène en établissant un lien entre les alternances syntaxiques caractéristiques de certains groupes de verbes et les composants sémantiques qui en sont responsables.

Dans la présente étude, nous cherchons à démontrer 1) que la classification automatique de verbes français selon leur comportement distributionnel aboutit souvent à des classes de verbes sémantiquement liés, 2) que les SSC extraits de corpus constituent une représentation fiable de la distribution syntaxique des verbes. Notre objectif est de mettre en place un algorithme de classification aussi général que possible, applicable à de nouveaux verbes, sous condition d'avoir une quantité suffisante d'occurrences dans le corpus pour créer un modèle fiable de leur distribution.

Notre approche est non supervisée, c'est-à-dire que nous ne fournissons au système aucune connaissance autre que les couples verbes-SSC issus de LEXSCHEM non filtré, avec les informations de fréquence relative qui ont ici une importance cruciale. Nous utilisons en entrée une version non filtrée de la ressource pour deux raisons : le filtrage peut être lui-même une source d'erreurs et, plus fondamentalement, la présence régulière de modifieurs particuliers avec certains verbes est un critère très pertinent pour le calcul des classes sémantiques.

Un léger filtrage est effectué lors de la première étape de la classification, dans le but de réduire l'espace de traits : les schémas qui ont moins de 5 occurrences parmi les verbes à classifier sont exclus. Le nombre de SSC différents — la taille de l'espace de traits — dépend ainsi du vocabulaire verbal utilisé dans l'expérimentation. Dans le cadre de cette expérimentation, nous avons travaillé avec un espace de traits composé de 433 SSC[1]. La représentation des verbes correspond à leur distribution sur tous les schémas considérés dans l'expérimentation (calculée par l'estimation du maximum de vraisemblance, à partir des données de LEXSCHEM) :

p(t|v) = f(v,t) / f(v)

où f(v) correspond à la fréquence du verbe, et f(v,t) à la fréquence du verbe avec le schéma.

Nous avons utilisé une méthode de regroupement (*clustering*) ascendante hiérarchique. Au début du processus, chaque verbe constitue un groupe à un seul élément (*cluster* — dans ce qui suit, on distingue la notion de « groupe », c'est-à-dire un regroupement obtenu automatiquement, de celle de « classe » correspondant à la référence élaborée manuellement). Lors de chaque itération, les deux groupes de verbes les plus similaires sont unifiés. Cette méthode produit un partitionnement, c'est-à-dire des groupes disjoints de manière à ce que chaque élément à classifier n'appartienne qu'à un seul groupe (*hard clustering*). Bien que cette approche ne permette pas de traiter la polysémie, nous l'avons tout de même choisie pour la facilité de l'interprétation qu'elle offre[2].

Les distributions ont été comparées avec trois mesures de similarité différentes :

- la divergence de Kullback-Leibler

$$D_{KL}(x\|y) = \sum_{i=1}^{n} x_i \cdot \log \frac{x_i}{y_i} \tag{F.1}$$

- la divergence de Jensen-Shannon

$$D_{JS}(x\|y) = \frac{1}{2} D_{KL}(x\|M) + \frac{1}{2} D_{KL}(y\|M) \tag{F.2}$$

où

$$M = \frac{1}{2}(x+y) \tag{F.3}$$

1. Puisque le lexique non filtré constitue l'entrée du processus, le nombre des SSC utilisés dépasse celui des SSC dans LEXSCHEM.

2. Une classification des verbes incorporant les problèmes de polysémie serait évidemment souhaitable mais ceci reste un problème ouvert pour le traitement des langues. Sur le plan pratique, un travail manuel reste nécessaire si l'on souhaite obtenir un résultat tout à fait fiable. Voir (Kipper et al., 2008) pour une expérience en ce sens, montrant les avantages d'une approche mixte, automatique puis manuelle.

– et la divergence oblique (*skew divergence*)

$$D_\alpha(x||y) = D_{KL}\left(x||\alpha y + (1-\alpha)\, x\right) \tag{F.4}$$

L'inconvénient de la divergence de Kullback-Leibler est de prendre une valeur indéfinie lorsque la probabilité *y(i)* est 0. Aussi une méthode de lissage simple a-t-elle été appliquée aux données : si la fréquence de cooccurrence du SSC avec le verbe dans le corpus égale zéro, cette valeur sera remplacée par 0.0001, donnant une estimation approximative de la fréquence relative (0.0001 / f(V), où f(V) est la fréquence observée du verbe). La divergence de Jensen-Shannon ainsi que la divergence oblique sont des variantes fondées sur la divergence de Kullback-Leibler qui évitent le problème des valeurs indéfinies par approximation de la valeur de la divergence de Kullback-Leibler. La divergence de Jensen-Shannon est la seule mesure symétrique — pour les autres mesures, le minimum de la distance a été considéré pour chaque paire de verbes comparés. La divergence oblique est une variante pondérée de la divergence de Kullback-Leibler proposée par (Lee, 2001). La pondération se fait par le paramètre libre a, dont la valeur optimale est proche de 1 : nous l'avons fixée à 0.99.

Aucune présupposition concernant le nombre et la cardinalité des groupes de verbes n'a été incorporée dans l'algorithme. Ainsi, le point d'arrêt du processus de classification dépend de deux paramètres : la distance maximale entre les centres des deux groupes à être unifiés et la cardinalité des groupes. Ce double paramétrage permet d'éviter l'effet de chaîne, c'est-à-dire le phénomène d'absorption de beaucoup de verbes par quelques groupes très nombreux. Des expérimentations ont été conduites avec des paramètres différents, et les valeurs optimales de la distance et de la cardinalité maximales ont été établies individuellement pour chaque mesure de distance lors des exécutions de test.

F.1.2 Evaluation

S. Schulte im Walde (Schulte im Walde, 2009) propose deux approches différentes pour évaluer une classification automatique :

1. mesurer la cohérence à l'intérieur des groupes de verbes obtenus par une mesure de similarité indépendante de celle utilisée pour la tâche de classification même,
2. comparer le résultat à une classification (manuelle) de référence.

Dans le cadre de notre expérimentation, nous visons à confirmer l'hypothèse qu'il existe un lien entre le comportement syntaxique des verbes et leurs propriétés sémantiques. Il ne suffit donc pas de démontrer que notre algorithme arrive à modéliser correctement les similarités distributionnelles entre les verbes : c'est la cohérence sémantique des classes qui doit être examinée par comparaison à la référence.

Pour ce faire, nous avons créé à la main une classification dite « de référence ». La référence est composée de 176 verbes, classés dans 16 classes différentes, qui ont d'abord été définies à partir de la classification de Levin, par traduction des verbes anglais. Pour assurer l'homogénéité des classes françaises et une certaine cohérence par rapport au travail de Levin, nous avons

vérifié que tous les verbes d'une même classe partageaient un certain nombre de constructions similaires fondamentales. Les classes de la référence sont donc caractérisées par un composant sémantique ainsi que par (au moins) une structure syntaxique en commun. Nous avons enfin vérifié la validité de ces classes en ayant recours au Lexique-Grammaire, et nous avons pu constater qu'en général les verbes d'une même classe se situent dans la même table du LG, à quelques exceptions près (ce qui montre le besoin d'une comparaison en profondeur des approches et des ressources — comparaison qui sort du cadre de cet article mais que nous menons en parallèle). Pour pouvoir évaluer l'extensibilité et la robustesse de la méthode, des verbes de fréquences différentes ont été inclus dans l'expérimentation. La cardinalité des classes varie entre 8 et 17.

Les résultats ont été évalués par rapport à la référence selon quatre mesures. La difficulté de la tâche de classification dépend du nombre de classes. Pour une classification à m classes, la valeur basse (*baseline*) de l'exactitude (*accuracy*) est de $1/m$, soit 0.0625 dans notre cas.

Mesure de distance	Card.	APP	mPURITY	ACC	F-measure
KL	6	0.13	0.48	0.30	0.36
KL	5	0.13	0.51	0.27	0.35
JS	4	0.21	0.60	0.28	0.39
JS	5	0.18	0.54	0.30	0.38
skew	4	0.22	0.62	0.27	0.37
skew	5	0.18	0.55	0.29	0.37
skew	9	0.16	0.47	0.35	0.40

Les groupes de verbes résultants ont été comparés à la référence d'une part par la mesure *Adjusted Pairwise Precision*, qui calcule la précision des groupes en prenant les verbes d'une même classe deux à deux (puis en comparant la sortie du système avec la référence pour vérifier s'ils appartiennent bien à la même classe dans les deux cas), et en prenant en compte la cardinalité (afin de pénaliser les petits groupes de verbes) :

$$APP(C) = \frac{1}{|C|}\sum_{i=1}^{C}\frac{paires\_correctes\_dans\_c_i}{paires\_dans\_c_i} \times \frac{c_i - 1}{c_i + 1} \tag{F.5}$$

D'autre part, il est possible d'associer les groupes de verbes résultant de la classification non supervisée aux classes de la référence en établissant la correspondance selon la classe sémantique prédominante à l'intérieur du groupe. Cela nous permet de calculer la pureté modifiée (*modified purity* — pureté moyenne des classes) et l'exactitude pondérée de classes (*weighted class accuracy* — rappel pondéré en fonction de la taille des classes dans la référence) (Korhonen et al., 2008). Lors du calcul de la pureté modifiée, les éléments qui n'appartiennent pas à la classe prédominante, ainsi que les singletons sont considérés comme des erreurs.

$$mPurity(C) = \frac{\sum_{n_{prevalent}(k_i)\geq 2} n_{prevalent}(k_i)}{|C|} \tag{F.6}$$

L'exactitude pondérée des classes peut être considérée comme une mesure de rappel : pour chaque classe de la référence, elle considère la quantité des verbes appartenant au groupe dominant associé à cette classe. Par définition, cette quantité ne peut pas dépasser la cardinalité maximale des groupes.

$$Acc\,(C) = \frac{\sum_{i=1}^{C} verbes\_dans\_GRP.DOM_i}{|C|} \tag{F.7}$$

La F-mesure a été calculée avec des poids égaux pour le rappel et la précision :

$$F = \frac{2 \times mPurity \times Acc}{mPurity + Acc} \tag{F.8}$$

En optimisant les paramètres pour la mesure APP, les groupes à quatre éléments donnent les meilleurs résultats. La précision forte semble soutenir le lien supposé entre les propriétés sémantiques et la distribution syntaxique observée dans le corpus, comme dans les exemples suivants :

```
groupe : errer voyager circuler naviguer
groupe : dire indiquer affirmer déclarer
groupe : signaler révéler montrer annoncer
groupe : ressentir définir désigner percevoir
groupe : rouspéter ronchonner grogner râler
```

Il est important de noter que cette qualité de la classification (20 % des classes sont parfaitement homogènes, 43 % contiennent 1 verbe incorrect au maximum) a été obtenue en utilisant une chaîne de traitement entièrement automatisée, de l'analyse de corpus jusqu'à la construction de l'espace de traits pour la classification. De plus, l'espace de traits est conçu pour être aussi général que possible, n'incorporant aucune connaissance préalable sur la classification de référence.

Cependant, les mesures de rappel pénalisent plus sérieusement la différence structurelle entre le regroupement résultant et la classification de référence, notamment en ce qui concerne le nombre et la cardinalité des groupes. La mesure d'exactitude montre que la cohérence des groupes baisse avec l'augmentation de la cardinalité au dessus de 4. En observant les résultats, nous pouvons noter que c'est souvent l'effet de chaîne qui affaiblit la cohérence des groupes de verbes : au lieu de réunir des verbes autour d'un composant sémantique central, ils sont composés d'une série de paires avec un lien sémantique qui se modifie constamment par l'ajout d'un nouvel élément. Par exemple, les groupes ci-dessous affichent une certaine cohérence sémantique, mais la relation sémantique est modifiée par rapport à la référence (les crochets indiquent la classification de référence) :

```
groupe : [resplendir pétiller scintiller] [vibrer]
groupe : [consterner ennuyer] [dévisager] [rosser]
groupe : [bougonner gémir] [trembler vaciller]
groupe : [grésiller geindre] [trembloter] [flamboyer]
groupe : [consolider renforcer] [réintégrer] [maintenir]
```

Plusieurs améliorations peuvent être envisagées. Une classification supervisée permettrait de spécifier les composants sémantiques centraux des classes, et ainsi d'adapter l'espace de traits (par filtrage ou pondération) à la tâche spécifique. Les schémas de sous-catégorisation les plus spécifiques aux classes donneront une idée des alternances qui caractérisent les classes sémantiques de verbes français. Le recours à d'autres traits (notamment des informations quant aux restrictions de sélection) permettrait d'obtenir des classes différentes et sans doute plus précises que celles obtenues en l'état.

F.2 Article publié dans les actes de la conférence CoLing en 2010

Cette section reproduit l'article :

Lin Sun, Thierry Poibeau, Anna Korhonen and Cédric Messiant. Investigating the cross-linguistic potential of VerbNet-style classification. In *Proceedings of CoLing*, 2010.

F.2.1 Introduction

A number of verb classifications have been built to support natural language processing (NLP) tasks (Grishman et al., 1994; Miller, 1995; Baker et al., 1998; Palmer et al., 2005; Kipper-Schuler, 2005; Hovy et al., 2006). These include both syntactic and semantic classifications, as well as ones which integrate aspects of both. Classifications which integrate a wide range of linguistic properties can be particularly useful for NLP applications suffering from data sparseness. One such classification is VerbNet (Kipper-Schuler, 2005).

Building on the well-known taxonomy of Levin (1993), VerbNet groups verbs (e.g. *deliver, post, dispatch*) into classes (e.g. SEND) on the basis of their shared meaning components and (morpho-)syntactic behaviour, identified in terms of meaning preserving (or extending) syntactic alternations called *diathesis alternations*. Such classes can be identified across the entire lexicon, and interestingly, they may also apply across languages, since the basic meaning components they are comprised of are said to be cross-linguistically applicable (Jackendoff, 1990).

Offering a powerful tool for generalization, abstraction and prediction, VerbNet classes have been used to support many important NLP tasks, including e.g. computational lexicography, parsing, word sense disambiguation, semantic role labeling, information extraction, question-answering, and machine translation (Swier et Stevenson, 2004; Dang, 2004; Shi et Mihalcea, 2005; Abend et al., 2008; Zapirain et al., 2008). However, to date their real-world exploitation has been limited because for most languages, no Levin-style classification is available.

Although manual classification is an option, it can be extremely costly (Kipper et al., 2008). In recent years, a variety of automatic approaches have been proposed which could be used to learn novel classifications in a cost-effective manner (Schulte im Walde, 2006; Joanis et al., 2008; Li et Brew, 2008; Ó Séaghdha et Copestake, 2008; Vlachos et al., 2009; Sun et Korhonen, 2009; Suzuki et Fukumoto, 2009) However, most work on Levin type classification has focussed on English. Large-scale research on other languages such as German (Schulte im Walde, 2006) and Japanese (Suzuki et Fukumoto, 2009) has focussed on semantic classification. Although there are similarities between the two classification systems, studies comparing the overlap between VerbNet and WordNet (Miller, 1995) have reported that the mapping is only partial and many to many due to fine-grained nature of classes based on synonymy (Kipper-Schuler, 2005; Shi et Mihalcea, 2005; Abend et al., 2008).

Only few studies have been conducted on Levin style classification for languages other than English. In their experiment involving 59 verbs and three classes, Merlo et al. (2002) applied a supervised approach developed for En-

glish to Italian, obtaining high accuracy (86.3%). In another experiment with 60 verbs and three classes, they showed that features extracted from Chinese translations of English verbs can improve English classification. These results are promising, but those from a later experiment by Ferrer (2004) are not. Ferrer applied a clustering approach developed for English to Spanish, and evaluated it against the manual classification of Vázquez et al. (2000), constructed using criteria similar (but not identical) to Levin's. This experiment involving 514 verbs and 31 classes produced results only slightly better than the random baseline.

In this paper, we investigate the cross-linguistic potential of Levin style classification further. In past years, verb classification techniques – in particular unsupervised ones – have improved considerably, making investigations for a new language more feasible. We take a recent verb clustering approach developed for English Sun et Korhonen (2009) and apply it to French – a major language for which no such experiment has been conducted yet. Basic NLP resources (corpora, taggers, parsers and subcategorization acquisition systems) are now sufficiently developed for this language for the application of a state-of-the-art verb clustering approach to be realistic.

Our investigation reveals similarities between the English and French classifications, supporting the linguistic hypothesis (Jackendoff, 1990) and the earlier result of Merlo et al. (2002) that Levin classes have a strong cross-linguistic basis. Not only the general methodology but also best performing features are transferable between the languages, making it possible to learn useful classes for French automatically and without the need for language-specific tuning.

F.2.2 French Verb Classes and the Gold Standard

The development of an automatic verb classification approach requires at least an initial gold standard. Some syntactic (Gross, 1975) and semantic (Vossen, 1998) verb classifications exist for French, along with ones which aim to integrate aspects of both (Saint-Dizier, 1998). Although such resources could be combined to hypothesise Levin-style classes for French (using e.g. an approach similar to that employed by Kipper et al. (2008)), we adopted a more direct approach : following the idea of Merlo et al. (2002), we translated a number of Levin classes from English to French.

We chose an English gold standard which has been used to evaluate several recent clustering works – that of Sun et al. (2008). It includes 17 fine-grained Levin classes. Each class has 12 member verbs whose predominant sense (according to the WordNet frequency data) belongs to that class. We evaluated each class in this resource as follows :

1. Member verbs were first translated to French. Where several relevant translations were identified, each of them was considered.
2. For each candidate verb, SCFs were identified and possible diathesis alternations were considered using the criteria of Levin (1993) : alternations must result in the same or extended verb sense. Only verbs sharing diathesis alternations were kept in the class – others were disrecarded.

For example, the gold standard class 31.1 AMUSE includes the following English verbs : *stimulate, threaten, shock, confuse, upset, overwhelm, scare, disappoint, delight, exhaust, intimidate* and *frighten*. Relevant French translations

Class No	Class	Verbs
9.1	PUT	accrocher, déposer, mettre, placer, répartir, réintégrer, empiler, emporter, enfermer, insérer, installer
10.1	REMOVE	ôter, enlever, retirer, supprimer, retrancher, débarrasser, soustraire, décompter, éliminer
11.1	SEND	envoyer, lancer, transmettre, adresser, porter, expédier, transporter, jeter, renvoyer, livrer
13.5.1	GET	acheter, prendre, saisir, réserver, conserver, garder, préserver, maintenir, retenir, louer, affréter
18.1	HIT	cogner, heurter, battre, frapper, fouetter, taper, rosser, brutaliser, éreinter, maltraiter, corriger,
22.2	AMALGAMATE	incorporer, associer, réunir, mélanger, mêler, unir, assembler, combiner, lier, fusionner
29.2	CHARACTERIZE	appréhender, concevoir, considérer, décrire, définir, dépeindre, désigner, envisager, identifier, montrer, percevoir, représenter, ressentir
30.3	PEER	regarder, écouter, examiner, considérer, voir, scruter, dévisager
31.1	AMUSE	abattre, accabler, briser, déprimer, consterner, anéantir, épuiser, exténuer, écraser, ennuyer, éreinter, inonder,
36.1	CORRESPOND	coopérer, participer, collaborer, concourir, contribuer, prendre part, s'associer, travaille
37.3	MANNER OF SPEAKING	râler, gronder, crier, ronchonner, grogner, bougonner, maugréer, rouspéter, grommeler, larmoyer, gémir, geindre, hurler, gueuler, brailler, chuchoter
37.7	SAY	dire, révéler, déclarer, signaler, indiquer, montrer, annoncer, répondre, affirmer, certifier, répliquer
43.1	LIGHT EMISSION	briller, étinceler, flamboyer, luire, resplendir, pétiller, rutiler, rayonner., scintiller
45.4	CHANGE OF STATE	mélanger, fusionner, consolider, renforcer, fortifier, adoucir, polir, atténuer, tempérer, pétrir, façonner, former
47.3	MODES OF BEING	trembler, frémir, osciller, vaciller, vibrer, tressaillir, frissonner, palpiter, grésiller, trembloter, palpiter
51.3.2	RUN	voyager, aller, se promener, errer, circuler, se déplacer, courir, bouger, naviguer, passer

TABLE F.1 – *A Levin style gold standard for French*

were identified for all of them : *abattre, accabler, briser, déprimer, consterner, anéantir, épuiser, exténuer, écraser, ennuyer, éreinter, inonder.* The majority of these verbs take similar SCFs and diathesis alternations, e.g. *Cette affaire écrase Marie (de chagrin), Marie est écrasée par le chagrin, Le chagrin écrase Marie* However, *stimuler* (*stimulate*) and *menacer* (*threaten*) do not, and they were therefore removed.

40% of translations were discarded from classes after step 2 was applied. The final version of the gold standard (shown in table F.1) includes 171 verbs in 16 classes. Each class is named according to the original Levin class. The smallest class (30.3) includes 7 verbs and the largest (37.3) 16. The average number of verbs per class is 10.7.

F.2.3 Verb Clustering

We performed an experiment where we
- took a French corpus and a SCF lexicon automatically extracted from that corpus using French NLP technology,
- extracted from these resources a range of features (lexical, syntactic and semantic) – a representative sample of those employed in recent English experiments (Joanis et al., 2008; Li et Brew, 2008; Ó Séaghdha et Copestake, 2008; Vlachos et al., 2009; Sun et Korhonen, 2009).
- clustered the features using a method which has proved promising in both English and German experiments : spectral clustering,
- evaluated the clusters both quantitatively (using the gold standard) and qualitatively,
- and finally, compared the performance of individual features to that recently obtained for English in order to gain a better understanding of the cross-linguistic and language-specific properties of verb classification

This work is described in the subsequent subsections.

Data : the LexSchem Lexicon

We extracted the features for clustering from LexSchem (Messiant et al., 2008). This large subcategorization lexicon provides SCF frequency information for 3,297 French verbs. It was acquired fully automatically from *Le Monde* newspaper corpus (200M words from the period 1991-2000) using ASSCI – a recent subcategorization acquisition system for French (Messiant, 2008).

Systems similar to ASSCI have been used in recent verb classification works e.g. (Schulte im Walde, 2006; Li et Brew, 2008; Ó Séaghdha et Copestake, 2008; Sun et Korhonen, 2009). Like these other systems, ASSCI takes raw corpus data as input. The data is first tagged and lemmatized using the Tree-Tagger and then parsed using the Syntex parser (Bourigault et al., 2005). Syntex is a shallow parser which employs a combination of statistics and heuristics to identify grammatical relations (GRs) in sentences.

ASSCI considers those GRs where the target verbs occur and constructs SCFs from nominal, prepositional and adjectival phrases, and infinitival and subordinate clauses. When a verb has no dependency, its SCF is considered as intransitive. Otherwise, ASSCI assumes no pre-defined list of SCFs but almost any combination of permitted constructions can appear as a candidate SCF. The number of automatically generated SCF types in LexSchem is 336.

Many of the candidate SCFs are noisy due to processing errors and the difficulty of argument-adjunct distinction. Most SCF systems operate on the basis of the assumption that true arguments occur in argument positions more frequently than adjuncts. Many of them also integrate sophisticated filters for removing noise from the system output. When LexSchem was evaluated using a relative frequency and heuristics -based filter its F-measure was 69 – which is similar to those of other current SCF systems (Messiant et al., 2008) However, we used the unfiltered version of LexSchem because previous work on English verb classification has showed that information about adjuncts can actually help verb clustering (Sun et al., 2008).

F.2.4 Features

Lexical entries in LexSchem provide a variety of material for verb clustering, including e.g. (statistical) information related to the part-of-speech (POS) tags, SCFs, argument heads, and adjuncts of verbs. Using this material, we constructed a range of features for experimentation. The first three include basic information about SCFs :

F1 : SCFs and their relative frequencies with individual verbs. SCFs abstract over particles and prepositions.

F2 : F1, with SCFs parameterized for the tense (the POS tag) of the verb.

F3 : F2, with SCFs parameterized for prepositions (PP).

The following six features include information about the lexical context (co-occurrences) of verbs. We adopt the best method of Li et Brew (2008) where collocations (COs) are extracted from the window of words immediately preceding and following a lemmatized verb. Stop words are removed prior to extraction.

F4, F6, F8 : COs are extracted from the window of 4, 6 and 8 words, respectively. The relative word position is ignored.

F5, F7, F9 : F4, F6 and F8 with the relative word position recorded.

The next four features include information about lexical preferences (LP) of verbs in argument head positions of specific GRs associated with the verb :

F10 : LP(PREP) : the type and frequency of prepositions in the preposition (PREP) relation.

F11 : LP(SUBJ) : the type and frequency of nouns in the subject (SUBJ) relation.

F12 : LP(IOBJ) : the type and frequency of nouns in the object (OBJ) and indirect object (IOBJ) relation.

F13 : LP(ALL) : the combination of F10-F13.

The final two features refine SCF features with LPs and semantic information about verb selectional preferences (SP) :

F14-F16 : F1-F3 parameterized for LPs.

F17 : F3 refined with SPs.

We adopt a fully unsupervised approach to SP acquisition using the method of Sun et Korhonen (2009), with the difference that we determine the optimal number of SP clusters automatically following Zelnik-Manor et Perona (2004).

The method is introduced in the following subsection. The approach involves (i) taking the GRs (subj, obj, iobj) associated with verbs, (ii) extracting all the argument heads in these GRs, and (iii) clustering the resulting N most frequent argument heads into M classes. The empirically determined N 200 was used. The method produced 40 SP clusters.

F.2.5 Clustering methods

Spectral clustering (SPEC) has proved promising in previous verb clustering experiments (Brew et Schulte im Walde, 2002; Sun et Korhonen, 2009) and other similar NLP tasks involving high dimensional feature space (Chen et al., 2006). Following Sun et Korhonen (2009) we used the MNCut spectral clustering (Meila et Shi, 2001) which has a wide applicability and a clear probabilistic interpretation (von Luxburg, 2007; Verma et Meila, 2005). However, we extended the method to determine the optimal number of clusters automatically using the technique proposed by (Zelnik-Manor et Perona, 2004).

Clustering groups a given set of verbs $V = \{v_n\}_{n=1}^{N}$ into a disjoint partition of K classes. SPEC takes a similarity matrix as input. All our features can be viewed as probabilistic distributions because the combination of different features is performed via parameterization. Thus we use the Jensen-Shannon divergence (JSD) to construct the similarity matrix. The JSD between two feature vectors v and v' is $d_{jsd}(v,v') = \frac{1}{2}D(v||m) + \frac{1}{2}D(v'||m)$ where D is the Kullback-Leibler divergence, and m is the average of the v and v'.

The similarity matrix W is constructed where $W_{ij} = \exp(-d_{jsd}(v,v'))$. In SPEC, the similarities W_{ij} are viewed as the connection weight ij of a graph G over V. The similarity matrix W is thus the adjacency matrix for G. The degree of a vertex i is $d_i = \sum_{j=1}^{N} w_{ij}$. A cut between two partitions A and A' is defined to be $\text{Cut}(A,A') = \sum_{m\in A, n\in A'} W_{mn}$.

The similarity matrix W is normalized into a stochastic matrix P.

$$P = D^{-1}W \tag{F.9}$$

The degree matrix D is a diagonal matrix where $D_{ii} = d_i$.

It was shown by Meila et Shi (2001) that if P has the K leading eigenvectors that are piecewise constant [3] with respect to a partition I^* and their eigenvalues are not zero, then I^* minimizes the multiway normalized cut(MNCut) :

$$\text{MNCut}(I) = K - \sum_{k=1}^{K} \frac{\text{Cut}(I_k, I_k)}{\text{Cut}(I_k, I)}$$

P_{mn} can be interpreted as the transition probability between vertices m, n. The criterion can thus be expressed as $\text{MNCut}(I) = \sum_{k=1}^{K}(1 - P(I_k \rightarrow I_k|I_k))$ (Meila, 2001), which is the sum of transition probabilities across different clusters. This criterion finds the partition where the random walks are most likely to happen within the same cluster. In practice, the leading eigenvectors of P are not piecewise constant. But we can extract the partition by finding the approximately equal elements in the eigenvectors using a clustering algorithm like K-MEANS.

As the value of K is not known beforehand, we use Zelnik-Manor et Perona (2004)'s method to estimate it. This method finds the optimal value by minimizing a cost function based on the eigenvector structure of W.

3. The eigenvector v is piecewise constant with respect to I if $v(i) = v(j) \forall i, j \in I_k$ and $k \in 1, 2...K$

Like Brew et Schulte im Walde (2002), we compare SPEC against a K-MEANS baseline. We used the Matlab implementation with euclidean distance as the distance measure.

F.2.6 Experimental evaluation

Data and pre-processing

Our initial plan was to experiment with all the 171 verbs in the gold standard (see Table F.1). However, we decided to exclude phrasal verbs (e.g. *faire disparaître*) and drop one class (40.2 NON-VERBAL EXPRESSION) which included reflexive verbs in French (e.g. *s'amuser, se moquer*) since multiword units would have been challenging for our method. Also verbs assigned to several classes due to polysemy were excluded. This left us with 147 verbs in 15 classes (10 verbs per class on average).

The SCF-based features (F1-F3 and F14-F17) were extracted directly from LexSchem. The CO (F4-F9) and LP features (F10-F13) were extracted from the raw and parsed corpus sentences, respectively, which were used for creating the lexicon. Features that only appeared once were removed. Feature vectors were normalized by the sum of the feature values before clustering. Since our clustering algorithms have an element of randomness, we repeated clustering multiple times. We report the results that minimize the distortion (the distance to cluster centroid).

Evaluation measures

We employ the same measures for evaluation as previously employed e.g. by Ó Séaghdha et Copestake (2008) and Sun et Korhonen (2009).

The first measure is modified purity (mPUR) – a global measure which evaluates the mean precision of clusters. Each cluster is associated with its prevalent class. The number of verbs in a cluster K that take this class is denoted by $n_{prevalent}(\mathrm{K})$. Verbs that do not take it are considered as errors. Clusters where $n_{prevalent}(\mathrm{K}) = 1$ are disregarded as not to introduce a bias towards singletons :

$$\mathrm{mPUR} = \frac{\sum_{n_{prevalent}(k_i)>2} n_{prevalent}(k_i)}{\text{number of verbs}}$$

The second measure is weighted class accuracy (ACC) : the proportion of members of dominant clusters DOM-CLUST$_i$ within all classes c_i.

$$\mathrm{ACC} = \frac{\sum_{i=1}^{C} \text{verbs in DOM-CLUST}_i}{\text{number of verbs}}$$

mPUR and ACC can be seen as a measure of precision(P) and recall(R) respectively. We calculate F measure as the harmonic mean of P and R :

$$\mathrm{F} = \frac{2 \cdot \mathrm{mPUR} \cdot \mathrm{ACC}}{\mathrm{mPUR} + \mathrm{ACC}}$$

The random baseline (BL) is calculated as follows :

$$\mathrm{BL} = 1/\text{number of classes}$$

F.2.7 Evaluation

Quantitative evaluation

In our first experiment, we evaluated 116 verbs – those which appeared in LexSchem the minimum of 150 times. We did this because English experiments

had shown that due to the Zipfian nature of SCF distributions, 150 corpus occurrences are typically needed to obtain a sufficient number of frames for clustering (Sun et al., 2008).

Table F.2 shows F-measure results for all the features. The 4th column of the table shows, for comparison, the results Sun et Korhonen (2009) obtained for English when they used the same features than us, clustered them using SPEC, and evaluated them against the English version of our gold standard, also using F-measure[4].

As expected, SPEC (the 2nd column) outperforms K-MEANS (the 3rd column) throughout the feature set. Looking at the basic SCF features F1-F3, we can see that they perform significantly better than the BL method. F3 performs the best among the three features both in French (50.6 F) and in English (63.3 F). We therefore use F3 as the SCF feature in F14-F17 (the same was done for English).

In French, most CO features (F4-F9) outperform SCF features. The best result is obtained with F7 : 55.1 F. This is clearly better than the best SCF result 50.6 (F3). This result is interesting since SCFs correspond better than COs with features used in manual Levin classification. Also, SCFs perform considerably better than COs in the English experiment (we only have the result for F4 available, but it is considerably lower than the result for F3). However, earlier English studies have reported contradictory results (e.g. Li et Brew (2008) showed that CO performs better than SCF in supervised verb classification), indicating that the role of CO features in verb classification requires further investigation.

Looking at the LP features, F13 produces the best F (52.7) for French which is slightly better than the best SCF result for the language. Also in English, F13 performs the best in this feature group and yields a higher result than the best SCF-based feature F3.

Parameterizing the best SCF feature F3 with LPs (F14-16) and SPs (F17) yields better performance in French. F15 and F17 have the F of 54.5 and 54.6, respectively. These results are so close to the result of the best CO feature F7 (55.1 – which is the highest result in this experiment) that the differences are not statistically significant. In English, the results of F14-F17 are similarly good; however, only F17 beats the already high performance of F13.

On the basis of this experiment, it is difficult to tell whether shallow CO features or more sophisticated SCF-based features are better for French. In the English experiment sophisticated features performed better, and the SCF-SP feature F17 was the best one. However, the English experiment employed a much larger dataset. These more sophisticated features may suffer from data sparseness in our French experiment since although we required the minimum of 150 occurrences per verb in LexSchem, verb clustering performance tends to improve when more data is available, and given the fine-grained nature of LexShem SCFs it is likely that more data is required for optimal performance.

We therefore performed another experiment with French on the full set of 147 verbs, using SPEC, where we investigated the effect of instance filtering on the performance of the best features from each feature group : F3, F7,

4. Note that the results for the two languages are not mutually comparable due to differences in test sets, data sizes, and feature extraction systems (see Section 8 for discussion). The results for English are included so that we can compare the relative performance of individual features in the two languages in question.

F13 and F17. The results shown in Table F.3 reveal that the performance of the features remains fairly similar until the instance threshold of 1000. When 2000 occurrences per verb are used, the differences become clearer, until at the threshold of 4000, it is obvious that the most sophisticated SCF-SP feature F17 is by far the best feature for French (65.4 F) and the SCF feature F3 the second best (60.5 F). The CO-feature F7 and the LP feature F13 are not nearly as good (53.4 and 51.0 F).

Although the results at different thresholds are not comparable due to the different number of verbs and classes (see columns 2-3), the results for features at the same threshold are. Those results suggest that when 2000 or more occurrences per verb are used, most features perform like they performed for English in the experiment of Sun et Korhonen (2009), with CO being the least informative[5] and SCF-SP being the most informative feature. The only exception is the LP feature which performed relatively better than CO in English.

		SPEC	K-MEANS	English
BL		6.7	6.7	6.7
F1	SCF	42.4	39.3	57.8
F2	SCF(POS)	45.9	40.3	46.7
F3	SCF(PP)	**50.6**	36.9	63.3
F4	CO(4)	50.3	38.2	40.9
F5	CO(4+loc)	48.8	26.3	-
F6	CO(6)	52.7	29.2	-
F7	CO(6+loc)	**55.1**	33.8	-
F8	CO(8)	54.2	36.4	-
F9	CO(8+loc)	54.6	37.2	-
F10	LP(PREP)	35.5	32.8	49.0
F11	LP(SUBJ)	33.7	23.6	-
F12	LP(OBJ)	50.1	33.3	-
F13	LP(ALL)	**52.7**	40.1	74.6
F14	SCF+LP(SUBJ)	50.3	40.1	71.7
F15	SCF+LP(OBJ)	**54.5**	35.6	74.0
F16	SCF+LP(SUBJ+OBJ)	53.4	36.2	73.0
F17	SCF+SP	54.6	39.8	80.4

TABLE F.2 – *Results for all the features for French (SPEC and K-means) and English (SPEC)*

THR	Verbs	Cls	F3	F7	F13	F17
0	147	15	43.7	57.5	43.3	50.1
50	137	15	47.9	56.1	44.8	49.1
100	125	15	49.2	54.3	44.8	49.5
150	116	15	50.6	55.1	52.7	54.6
200	110	15	54.9	52.9	49.7	52.5
400	96	15	52.7	52.9	43.9	53.2
1000	71	15	51.4	54.0	44.8	54.5
2000	59	12	52.3	45.9	42.7	53.5
3000	51	12	55.7	49.0	46.8	59.2
4000	43	10	60.5	53.4	51.0	**65.4**

TABLE F.3 – *The effect of verb frequency on performance*

5. However, it is worth noting that CO is not a useless feature. As table F.3 shows, when 150 or fewer occurrences are available for a verb, CO outperforms all the other features in French, compensating for data sparseness.

Qualitative evaluation

We conducted qualitative analysis of the clusters for French, focusing on those created using SPEC with F17 and F3.

Verbs in the gold standard classes 29.2, 36.1, 37.3, 37.7 and 47.3 (Table F.1) performed particularly well, with the majority of member verbs found in the same cluster. These verbs are ideal for clustering because they have distinctive syntactic-semantic characteristics. For example, verbs in 29.2 CHARACTERIZE class (e.g. *concevoir, considérer, dépeindre*) not only have a very specific meaning but they also take high frequency SCFs involving the preposition *comme* (Eng. *as*) which is not typical to many other classes. Interestingly, Levin classes 29.2, 36.1, 37.3, and 37.7 were among the best performing classes also in the English supervised verb classification experiment of Sun et al. (2008) (which employed the English version of our gold standard) because these classes have distinctive characteristics also in English.

The benefit of sophisticated features which integrate also semantic (SP) information (F17) is particularly evident for classes with non-distinctive syntactic charateristics. For example, the intransitive verbs in 43.1 LIGHT EMISSION class (e.g. *briller, étinceler, flamboyer*) are difficult to cluster based on syntax only, but semantic features work because the verbs pose strong SPs on their subjects (entities capable of light emission). In the experiment of Sun et al. (2008), 43.1 was the worst performing class for English, possibly because no semantic features were used in the experiment.

The most frequent source of error is syntactic idiosyncracy. This is particularly evident for classes 10.1 REMOVE and 45.4 CHANGE OF STATE. Although verbs in these classes can take similar SCFs and alternations, only some of them are frequent in data. For example, the SCF *ôter X à Y* is frequent for verbs in 10.1, but not *ôter X de Y*. Although class 10.1 did not suffer from this problem in the English experiment of Sun et al. (2008), class 45.4 did. Class 45.4 performs particularly bad in French also because its member verbs are low in frequency.

Some errors are due to polysemy, caused partly by the fact that the French version of the gold standard was not controlled for this factor. Some verbs have their predominant senses in classes which are missing in the gold standard, e.g. the most frequent sense of *retenir* is *memorize*, not *keep* as in the gold standard class 13.5.1. GET.

Finally, some errors are not true errors but demonstrate the capability of clustering to learn novel information. For example, the CHANGE OF STATE class 45.4 includes many antonyms (e.g. *weaken* vs. *strenghten*). Clustering (using F17) separates these antonyms, so that verbs *adoucir, atténuer* and *tempérer* appear in one cluster and *consolider* and *renforcer* in another. Although these verbs share the same alternations, their SPs are different. For the same reason, verbs in LIGHT EMISSION class 43.1 end up in different clusters, depending on whether they describe abstract or concrete light emission.

The opposite effect can be observed when clustering maps together classes which are actually semantically and syntactically related (e.g. 36.1 CORRESPOND and 37.7 SPEAK). Such classes are distinct in Levin and VerbNet, because these resources do not to draw links between semantically similar classes belonging to different main classes.

Cases such as these show the potential of clustering in discovering novel

valuable information in data. It is encouraging that we have observed this effect in this first clustering experiment in French.

F.2.8 Discussion and Conclusion

We have seen that when sufficient corpus data is available, there is a strong correlation between the types of features which perform the best in English and French. Interestingly, we have also seen that when the best features are used, many individual Levin classes have similar performance in the two languages.

Due to differences in language-specific data sets and sizes direct comparison of the actual performance figures for English and French is not possible. When considering the general level of performance, our best performance for French (65.4 F) is clearly lower than the best performance for English in the experiment of Sun et Korhonen (2009). However, it compares favourably to the performance of other state-of-the-art (even supervised) systems for English verb classification (Joanis et al., 2008; Li et Brew, 2008; Ó Séaghdha et Copestake, 2008; Vlachos et al., 2009). This is impressive considering that we experimented with a fully unsupervised method originally developed for another language.

Our experiment suggests that when aiming to improve performance further, employing larger data is critical. Most recent experiments on English have employed bigger test and data sets, and unlike us, some of them have only considered the predominant senses of medium-high frequency verbs (Ó Séaghdha et Copestake, 2008; Vlachos et al., 2009; Sun et Korhonen, 2009). As seen in subsection F.2.7, such differences in data can have significant impact on performance.

However, parser and feature extraction performance can also play a big role in overall accuracy, and should therefore be investigated further. When Sun et Korhonen (2009) evaluated their basic SCF feature (equivalent to F1) using the same corpus data and gold standard but an older version of the parser and the SCF extraction system, the F dropped dramatically : from 57.8 to 38.3. The relatively low performance of basic LP features in French suggests that at least some of the current errors are due to parsing. Future research should therefore investigate the source of error at different stages of processing.

In the future, it would also be interesting to investigate whether performance on French can be further enhanced by language-specific tuning (e.g. by experimenting with language specific features such as auxiliary classes).

Methodology similar to ours has yielded promising results on semantic verb classification in German (Schulte im Walde, 2006) and Japanese (Suzuki et Fukumoto, 2009). However, these studies have not focussed on Levin style classes, and have not explored cross-linguistic transfer. The works most related to ours are those of Merlo et al. (2002) and Ferrer (2004). Our results contrast with those of Ferrer who showed that a clustering approach does not transfer well from English to Spanish. However, her experiment used basic SCF and named entity features only, and a clustering algorithm less suitable for high dimensional data.

Like us, Merlo et al. (2002) created a gold standard by translating Levin classes to another language (Italian). They also applied a classification approach developed for English to Italian, and reported good overall performance using features developed for English. Although the experiment was very small

in scale (involving three classes and a few features only), and although it involved a use of a supervised classification technique, the results are in agreement with our results from this larger, unsupervised experiment with French.

These experiments further support the linguistic hypothesis that Levin-style classification can be cross-linguistically applicable or overlapping (Levin, 1993). A clustering technique such as the one presented here could be used as a helpful tool to investigate this hypothesis further, and to find out whether classifications are similar across a wider range of more diverse languages. From the NLP perspective, the fact that an unsupervised technique developed for one language can be applied to another language without substantial changes in the methodology means that automatic techniques can be used to hypothesise useful Levin-style classes in a cost-effective manner Kipper et al. (2008). This, in turn, can facilitate the creation of VerbNets for new languages.

Liste des publications liées à la thèse

G

Articles de revue

Cédric Messiant, Kata Gábor et Thierry Poibeau. Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français. *Traitement Automatique des Langues*, 2010.

Cédric Messiant et Thierry Poibeau. Automatic Lexical Acquisition from Corpora, some Limitations and some Tentative Solutions. *Cahiers du Cental (numéro spécial "eLexicography in the 21st Century : New Challenges, New Applications")*, Presses Universitaires de Louvain, 2010.

Communications internationales avec actes

Lin Sun, Thierry Poibeau, Anna Korhonen and Cédric Messiant. Investigating the cross-linguistic potential of VerbNet-style classification. In *Proceedings of CoLing*, 2010.

Cédric Messiant. 2008. ASSCI : A Subcategorization Frames Acquisition System for French Verbs. In *Proceedings of the Association for Computational Linguistics (ACL, Student Research Workshop)*, Columbus, Ohio. pp. 55–60.

Cédric Messiant, Anna Korhonen, and Thierry Poibeau. 2008. LexSchem : A Large Subcategorization Lexicon for French Verbs. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marrakech.

Thierry Poibeau and Cédric Messiant. 2008. Do we still need gold standard for evaluation ? In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marrakech.

Communications nationales avec actes

Cédric Messiant, Takuya Nakamura et Stavroula Voyatzi. 2009. La complémentarité des approches manuelle et automatique en acquisition lexicale. Dans *Actes de la 16ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Senlis.

Cédric Messiant et Thierry Poibeau. 2009. Acquisition d'informations lexicales à partir de corpus. Dans *Troisième colloque international de l'Association Française de Linguistique Cognitive (AFLICO)*, Nanterre.

BIBLIOGRAPHIE

Anne Abeillé, Lionel Clément, et François Toussenel. Building a treebank for french. Dans Anne Abeillé, éditeur, *Treebanks : Building and Using Parsed Corpora*, pages 165–187, Dordrecht, 2003. Kluwer Academic Publishers. (Cité page 37.)

Anne Abeillé. *Les nouvelles syntaxes*. Armand Colin, Paris, 1993. (Cité page 1.)

Omri Abend, Roi Reichart, et Ari Rappoport. A Supervised Algorithm for Verb Disambiguation into VerbNet Classes. Dans *Proceedings of CoLing*, pages 9–16, 2008. (Cité pages 115 et 144.)

Steven Abney et Marc Light. Hiding a Semantic Class Hierarchy in a Markov Model. Dans A. Kehler et A. Stolcke, éditeurs, *Proceedings of ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 1–8, College Park, Maryland, 1999. (Cité page 124.)

Eneko Agirre et Philip Edmonds, éditeurs. *Word Sense Disambiguation : Algorithms and Applications*. Springer, Berlin, 2007. (Cité pages 2 et 17.)

Afra Alishahi et Suzanne Stevenson. A cognitive model for the representation and acquisition of verb selectional preferences. Dans *ACL Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 41–48, Prague, Czech Republic, 2007. (Cité page 116.)

Collin F. Baker, Charles J. Fillmore, et John B. Lowe. The Berkeley FrameNet Project. Dans *Proceedings of COLING-ACL*, pages 86–90, 1998. (Cité pages 14, 114 et 144.)

Mark C. Baker. Thematic Roles and Syntactic Structure. *Elements of Grammar*, pages 73–137, 1997. (Cité page 14.)

Somnath Banerjee, Dipankar Das, et Sivaji Bandyopadhyay. Bengali verb subcategorization frame acquisition : a baseline model. Dans *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, pages 76–83, Morristown, NJ, USA, 2009. Association for Computational Linguistics. (Cité page 54.)

Claire Blanche-Benveniste, José Deulofeu, Jean Stéfanini, et Karel van den Eynde. *Pronom et syntaxe. L'approche pronominale et son application au français*. SELAF, Paris, 1984. (Cité page 31.)

Jean-Paul Boons, Alain Guillet, et Christian Lecrère. *La structure des phrases simples en français*. Droz, Genève, 1976. (Cité page 3.)

Andrée Borillo. Remarques sur les verbes symétriques du français. *Langue française*, 11 :17–31, 1971. (Cité page 3.)

Didier Bourigault. *Un analyseur syntaxique opérationnel : SYNTEX*. Habilitation à Diriger des Recherches, Université Toulouse 2, Toulouse, 2007. (Cité page 62.)

Didier Bourigault, Marie-Paule Jacques, Cécile Fabre, Cécile Frérot, et Sylwia Ozdowska. Syntex, analyseur syntaxique de corpus. Dans *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan, 2005. (Cité pages 60, 62, 103 et 147.)

Michael R. Brent. Automatic acquisition of subcategorization frames from untagged text. Dans *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 209–214, Berkeley, CA, 1991. (Cité pages 2 et 49.)

Michael R. Brent. From grammar to lexicon : Unsupervised learning of lexical syntax. *Computational Linguistics*, 19 :203–222, 1993. (Cité pages vi, 49, 53 et 79.)

Joan Bresnan et Annie Zaenen. Deep unaccusativity in lfg. Dans K. Dziwirek, éditeur, *Grammatical Relations. A Cross-Theoretical Perspective*. Center for the Study of Language and Information, Stanford University, 1990. (Cité page 1.)

Chris Brew et Sabine Schulte im Walde. Spectral clustering for german verbs. Dans *Proceedings of EMNLP*, pages 117–124, 2002. (Cité pages 149 et 150.)

Ted Briscoe et John Carroll. Automatic extraction of subcategorization from corpora. Dans *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC, 1997. (Cité pages 51, 52, 53 et 79.)

John Carroll et Ted Briscoe. Apportioning development effort in a probabilistic lr parsing system through evaluation. Dans *Proceedings of the ACL/SIGDAT Conference on Empirical Methods in Natural Language Processing*, pages 92–100, University of Pennsylvania, PA, 1996. (Cité page 52.)

John Carroll, Guido Minnen, et Ted Briscoe. Can subcategorisation probabilities help a statistical parser ? Dans *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*, Montreal (Canada), 1998. (Cité pages 7, 23, 53 et 125.)

Y. Kiliçaslan, E. Uzun, H.V. Agun, et E. Uçar. Automatic Acquisition of Subcategorization Frames for Turkish with Purely Statistical Methods. Dans *Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, INISTA-2007*, pages 11–15, Istanbul, Turkey, 2007. (Cité page 54.)

Jinxiu Chen, Dong-Hong Ji, Chew Lim Tan, et Zheng-Yu Niu. Unsupervised relation disambiguation using spectral clustering. Dans *Proceedings of COLING/ACL*, pages 89–96, 2006. (Cité page 149.)

Paula Chesley et Susanne Salmon-Alt. Automatic extraction of subcategorization frames for french. Dans *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Genua (Italy), 2006. (Cité pages 4, 54, 82 et 124.)

Stephen Clark et David J. Weir. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2) :187–206, 2002. (Cité page 124.)

Lionel Clément, Benoît Sagot, et Bernard Lang. Morphology Based Automatic Acquisition of Large-coverage Lexica. Dans *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1841–1844, may 2004. (Cité page 34.)

Matthieu Constant et Elsa Tolone. A Generic Tool to Generate a Lexicon for nlp from Lexicon-Grammar Tables. Dans Michele De Gioia, éditeur, *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008). Seconde partie.* Aracne, 2010. (Cité pages 29 et 45.)

Ann Copestake. *The Representation of Lexical Semantic Information.* PhD thesis, University of Sussex, 1992. (Cité page 1.)

Alan D. Cruse. *Lexical semantics.* Cambridge University Press, Cambridge, 1986. (Cité pages 2 et 14.)

Hoa Trang Dang. *Investigations into the Role of Lexical Semantics in Word Sense Disambiguation.* PhD thesis, CIS, University of Pennsylvania, 2004. (Cité pages 115 et 144.)

Laurence Danlos. Les lexiques en traitement automatique du langage naturel. Dans *Proceedings of the 3rd meeting on Langage Industry*, Grossetto, 1988. (Cité page 1.)

Laurence Danlos et Benoît Sagot. Constructions pronominales dans dicovalence et le lexique-grammaire – intégration dans le lefff. Dans *Actes du Colloque Lexique et Grammaire*, L'Aquila, Italie, 2008. (Cité page 34.)

Tim Van de Cruys et Begoña Villada Moirón. Lexico-semantic multiword expression extraction. Dans Peter Dirix et al., éditeur, *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN)*, pages 175–190, University of Leuven, Leuven, Belgium, 2007. (Cité page 90.)

Jacques Dendien et Jean-Marie Pierrel. Le trésor de la langue française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues*, 44 (2), 2003. (Cité page 36.)

David Dowty. Thematic proto-roles and argument selection. *Language*, 67 (3) :547–619, 1991. (Cité page 15.)

Jean Dubois et Françoise Dubois-Charlier. *Les verbes français.* Larousse, Paris, 1997. (Cité pages 34 et 36.)

Mireille Ducassé et Sébastien Ferré. Aide à la décision multicritère : cohérence et équité grâce à l'analyse de concepts. Dans *Modèles et Apprentissage en Sciences Humaines et Sociales*, 2009. (Cité page 116.)

Cécile Fabre et Didier Bourigault. Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. *Journal of French Language Studies*, 18(1) :87–102, 2008. (Cité pages 90 et 122.)

Ingrid Falk. *Création automatique de classes sémantiques verbales pour le français.* Mémoire de Master, LORIA, Nancy, 2008. (Cité page 116.)

Christiane Fellbaum, éditeur. *WordNet : An Electronic Lexical Database.* Language, Speech, and Communication. MIT Press, Cambrige, MA, May 1998. (Cité pages 16 et 124.)

Eva Esteve Ferrer. Towards a semantic classification of spanish verbs based on subcategorisation information. Dans *Proceedings of ACL Student Research Workshop*, 2004. (Cité pages 145 et 154.)

Charles J. Fillmore. The grammar of hitting and breaking. Dans R. Jacobs et P. Rosenbaum, éditeurs, *Readings in English Transformational Grammar*, pages 120–133. 1967. (Cité page 114.)

Charles J. Fillmore. The case for case. Dans Emmon Bach et Robert T. Harms, éditeurs, *Universals in linguistic theory*, pages 1–88. Holt, Rinehart and Winston, 1968. (Cité page 13.)

Karen Fort et Bruno Guillaume. PrepLex : un lexique des prépositions du français pour l'analyse syntaxique. Dans *Actes de la 14ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Toulouse France, 2007. (Cité pages 74 et 76.)

Gilles Francopoulo. Tagparser et technolangue-easy. Dans *Actes de l'Atelier technolangue, TALN'05*, Dourdan, 2005. (Cité page 55.)

Nuria Gala et André Valli. Building a computational lexicon of verbal syntactic constructions in french. Dans *Proceedings of PAPILLON-05, Workshop on Multilingual Lexical Databases*, Chiang Rai, Thailande, 2005. (Cité page 36.)

Claire Gardent. Evaluating an automatically extracted lexicon. Dans *Proceedings of the 4th Language & Technology Conference*, Poznan, Poland, 2009. (Cité pages 55 et 79.)

Claire Gardent, Bruno Guillaume, Guy Perrier, et Ingrid Falk. Extraction d'information de sous-catégorisation à partir des tables du ladl. Dans *Actes de Traitement Automatique des Langues Naturelles*, Louvain, Belgique, 2006. (Cité pages 28 et 45.)

Claire Gardent et Alejandra Lorenzo. Identifying Sources of Weakness in Syntactic Lexicon Extraction. Dans *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). (Cité pages 4, 55, 96 et 122.)

Daniel Gildea et Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3) :245–288, 2002. (Cité page 124.)

Genevieve Gorrell. *Acquiring subcategorisation from textual corpora.* M. Phil. thesis, University of Cambridge, UK, 1999. (Cité pages 53 et 79.)

Ralph Grishman, Catherine Macleod, et Adam Meyers. Comlex Syntax : Building a Computational Lexicon. Dans *Proceedings of the International Conference on Computational Linguistics*, pages 268–272, Kyoto, Japan, 1994. (Cité pages 114 et 144.)

Gaston Gross. *Les expressions figées en français : noms composés et autres locutions*. Ophrys, Paris, 1996. (Cité page 21.)

Maurice Gross. *Méthodes en syntaxe*. Hermann, Paris, 1975. (Cité pages 1, 2, 3, 10, 11, 16, 21, 24, 36, 114 et 145.)

Maurice Gross. Les bases empiriques de la notion de prédicat sémantique. *Langages*, 63 :7–52, 1981. (Cité page 21.)

Maurice Gross. Les limites de la phrase figée. *Langages*, 90 :7–22, 1988. (Cité pages 1 et 21.)

Maurice Gross. Constructing lexicon-grammars. *Computational Approaches to the Lexicon*, pages 213–263, 1994. (Cité page 24.)

Maurice Gross. La fonction sémantique des verbes supports. *Travaux de linguistique*, 37 :25–46, 1998. (Cité page 21.)

Jeffrey S. Gruber. *Studies in lexical relations*. PhD thesis, MIT, 1965. (Cité page 13.)

Xiwu Han, Tiejun Zhao, Haoliang Qi, et Hao Yu. Subcategorization acquisition and evaluation for chinese verbs. Dans *Proceedings of the 20th international conference on Computational Linguistics (CoLing)*, page 723, Morristown, NJ, USA, 2004. Association for Computational Linguistics. (Cité page 54.)

Zellig Harris. *Notes du cours de syntaxe*. Seuil, Paris, 1976. (Cité page 24.)

Eduard Hovy, Mitch Marcus, Martha Palmer, L. Ramshaw, et R. Weischedel. OntoNotes : The 90% Solution. Dans *HLT/NAACL*, 2006. (Cité pages 114 et 144.)

Ray S. Jackendoff. *Semantics and Cognition*. MIT Press, Cambridge/MA, 1972. (Cité page 13.)

Ray S. Jackendoff. *Semantic Structures*. The MIT Press, Cambridge, Massachusetts, 1990. (Cité pages 2, 14, 144 et 145.)

Eric Joanis, Suzanne Stevenson, et David James. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3) :337–367, 2008. (Cité pages 115, 144, 147 et 154.)

J.G Kalbfleisch. *Probability and Statistical Inference, Volume 2 : Statistical Inference*. Springer-Verlag, New York, 1985. (Cité page 49.)

Karin Kipper, Anna Korhonen, Neville Ryant, et Martha Palmer. A large-scale classification of english verbs. *Journal of Language Resources and Evaluation*, 42(1) :21–40, 2008. (Cité pages 3, 115, 140, 144, 145 et 155.)

Karin Kipper-Schuler. *VerbNet : A Broad-coverage, Comprehensive Verb Lexicon*. University of Pennsylvania, PA, 2005. (Cité pages 14, 114 et 144.)

Philipp Koehn. Europarl : A parallel corpus for statistical machine translation. Dans *MT Summit*, Phuket Island, Thailand, 2005. (Cité page 105.)

Jean-Pierre Koenig et Anthony Davis. Semantically transparent linking in hpsg. Dans *Proceedings of the HPSG03 Conference*, pages 222–235, East Lansing, Michigan, 2000. (Cité page 1.)

Anna Korhonen. Using semantically motivated estimates to help subcategorization acquisition. Dans *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 216–223, Hong Kong, 2000. (Cité pages 51, 53, 115, 122 et 124.)

Anna Korhonen. *Subcategorization acquisition*. PhD thesis, University of Cambridge, 2002. (Cité pages 12 et 51.)

Anna Korhonen et Ted Briscoe. Extended lexical-semantic classification of english verbs. Dans Dan Moldovan et Roxana Girju, éditeurs, *Proceedings of the HLT-NAACL 2004 : Workshop on Computational Lexical Semantics*, pages 38–45, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. (Cité page 115.)

Anna Korhonen, G. Gorrell, et D. McCarthy. Statistical filtering and subcategorization frame acquisition. Dans *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 199–205, Hong Kong, 2000. (Cité pages 51, 53, 78, 82, 99 et 124.)

Anna Korhonen, Yuval Krymolowski, et Ted Briscoe. A large subcategorization lexicon for natural language processing applications. Dans *Proceedings of the 5th international conference on Language Resources and Evaluation*, Genova, Italy, 2006. (Cité pages 51, 53 et 82.)

Anna Korhonen, Yuval Krymolowski, et Nigel Collier. The choice of features for classification of verbs in biomedical texts. Dans *Proceedings of the 2008 COLING conference*, pages 449–456, Manchester, 2008. (Cité page 142.)

Anna Korhonen, Yuval Krymolowski, et Zvika Marx. Clustering polysemic subcategorization frame distributions semantically. Dans *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, pages 64–71, Sapporo, 2003. (Cité pages 115 et 116.)

Anna Kupść. Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. Dans *Actes de la 14ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Toulouse, June 2007. (Cité page 37.)

Anna Kupść. Adjectives in treelex. Dans S. Wierzchoń et K. Trojanowski M. Klopotek, A. Przepiórkowski, éditeur, *Proceedings of the 16th International Conference Intelligent Information Systems*, pages 287–296, Zakopane, Poland, 2008. Academic Publishing House EXIT. (Cité page 38.)

Anna Kupść et Anne Abeillé. Growing treelex. Dans A. Gelbukh, éditeur, *Proceedings of the 9th International Conference, CICLing 2008*, pages 28–39, Haifa, Israel, 2008a. Lecture Notes in Computational Linguistics no 4919. (Cité page 37.)

Anna Kupść et Anne Abeillé. Treelex : A subcategorisation lexicon for french verbs. Dans *Proceedings of the First International Conference on Global Interoperability for Language Resources*, Hong Kong, 2008b. (Cité pages 37 et 95.)

Eric Laporte. Mots et niveau lexical. Dans Jean-Marie Pierrel, éditeur, *Ingénierie des langues*, pages 25–49. Hermés, 2000. (Cité page 1.)

Christian Leclère. The lexicon-grammar of french verbs : a syntactic database. *Linguistic Informatics - State of the Art and the Future*, pages 29–45, 2005. (Cité pages 24 et 26.)

Lilian Lee. On the effectiveness of the skew divergence for statistical language analysis. Dans *Proceedings of the Workshop on Artificial Intelligence and Statistics*, Key west, Florida, 2001. (Cité page 141.)

Alessandro Lenci, Barbara McGillivray, Simonetta Montemagni, et Vito Pirrelli. Unsupervised Acquisition of Verb Subcategorization Frames from Shallow-Parsed Corpora. Dans *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). (Cité page 54.)

Beth Levin. *English Verb Classes and Alternations : a preliminary investigation.* University of Chicago Press, Chicago and London, 1993. (Cité pages 1, 2, 3, 16, 53, 114, 139 et 144.)

Beth Levin et Malka Rappaport Hovav. *Argument Realization.* Cambridge university Press, Cambridge, 2005. (Cité page 3.)

Hang Li et Naoki Abe. Generalizing case frames using a thesaurus and the mdl principle. *Computational Linguistics*, 24 :239–248, 1998. (Cité page 124.)

Jianguo Li et Chris Brew. Which are the best features for automatic verb classification. Dans *Proceedings of the Meeting of the Association for Computational Linguistics (ACL–HLT)*, pages 434–442, Columbus, Ohio, 2008. (Cité pages 115, 116, 144, 147, 148, 151 et 154.)

John Lyons. *Introduction to Theoretical Linguistics.* Cambridge University Press, Cambridge, 1968. (Cité pages 10 et 11.)

Christopher D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. Dans *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 235–242, Columbus, Ohio, 1993. (Cité pages 2, 50, 51, 53, 55 et 79.)

Christopher D. Manning. Probabilistic syntax. Dans MIT Press, éditeur, *Probabilistic Linguistics*, pages 289–341. R. Bod, J. Hay, S. Jannedy, 2003. (Cité pages 19 et 77.)

M. Maragoudakis, K. Kermanidis, N. Fakotakis, et G. Kokkinakis. Learning Automatic Acquisition of Subcategorization Frames Using Bayesian Inference and Support Vector Machines. *Data Mining, IEEE International Conference on*, 0 :623, 2001. (Cité page 54.)

Svetoslav Marinov. Automatic Extraction of Subcategorization Frames for Bulgarian. Dans *Proceedings of the Ninth ESSLLI Student Session*, Edinburgh, 2004. (Cité page 37.)

Diana McCarthy. *Lexical Acquisition at the Syntax-Semantics Interface : Diathesis Alternations.* PhD Thesis, University of Sussex, 2001. (Cité pages 2, 17 et 115.)

Marina. Meila. The multicut lemma. Rapport technique, University of Washington, 2001. (Cité page 149.)

Marina Meila et Jianbo Shi. A random walks view of spectral segmentation. Dans *AISTATS*, 2001. (Cité pages 117 et 149.)

Igor Mel'cuk. Verbes supports sans peine. *Lingvisticae Investigationes*, 27(2) : 203–217, 2004. (Cité page 21.)

Paola Merlo, Suzanne Stevenson, Vivian Tsang, et Gianluca Allaria. A multilingual paradigm for automatic verb classification. Dans *Proceedings of ACL*, 2002. (Cité pages 144, 145 et 154.)

Piet Mertens. Restrictions de sélection et réalisations syntagmatiques dans dicovalence. conversion vers un format utilisable en tal. Dans *Actes de la 17ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Montreal, Canada, juillet 2010. (Cité page 96.)

Cédric Messiant, Anna Korhonen, et Thierry Poibeau. LexSchem : A Large Subcategorization Lexicon for French Verbs. Dans *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marrakech, 2008. (Cité pages 147 et 148.)

Cédric Messiant. A Subcategorization Acquisition System for French Verbs. Dans *Proceedings of the ACL-08 : HLT Student Research Workshop*, pages 55–60, Columbus, Ohio, June 2008. Association for Computational Linguistics. (Cité pages 82 et 147.)

Cédric Messiant, Kata Gábor, et Thierry Poibeau. Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français. *Traitement Automatique des Langues*, 2010. (Cité pages 117 et 120.)

Cédric Messiant, Takuya Nakamura, et Stavroula Voyatzi. La complémentarité des approches manuelle et automatique en acquisition lexicale. Dans *Actes de la 16ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Senlis, 2009. (Cité pages 26 et 27.)

George A. Miller. Wordnet : a lexical database for english. *Communications of the ACM*, 1995. (Cité pages 16, 114 et 144.)

Philip Miller. Compléments et circonstants : distinction syntaxique ou sémantique ? Dans J.-C.Souesme, éditeur, *Actes de l'Atelier de Linguistique. SAES 1997. Cycnos*, volume 15, Nice, 1997. (Cité page 11.)

Erwann Moreau, Isabelle Tellier, Antonio Balvet, Grégoire Laurence, Antoine Rozenknop, et Thierry Poibeau. Annotation fonctionnelle de corpus arborés avec des champs aléatoires conditionnels. Dans *Actes de la 16ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Senlis, 2009. (Cité page 2.)

Diana McCarthy. *Lexical Acquisition at the Syntax-Semantics Interface : Diathesis Alternations*. PhD Thesis, University of Sussex, 2001. (Cité pages 2, 17 et 115.)

Marina. Meila. The multicut lemma. Rapport technique, University of Washington, 2001. (Cité page 149.)

Marina Meila et Jianbo Shi. A random walks view of spectral segmentation. Dans *AISTATS*, 2001. (Cité pages 117 et 149.)

Igor Mel'cuk. Verbes supports sans peine. *Lingvisticae Investigationes*, 27(2) : 203–217, 2004. (Cité page 21.)

Paola Merlo, Suzanne Stevenson, Vivian Tsang, et Gianluca Allaria. A multilingual paradigm for automatic verb classification. Dans *Proceedings of ACL*, 2002. (Cité pages 144, 145 et 154.)

Piet Mertens. Restrictions de sélection et réalisations syntagmatiques dans dicovalence. conversion vers un format utilisable en tal. Dans *Actes de la 17ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Montreal, Canada, juillet 2010. (Cité page 96.)

Cédric Messiant, Anna Korhonen, et Thierry Poibeau. LexSchem : A Large Subcategorization Lexicon for French Verbs. Dans *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marrakech, 2008. (Cité pages 147 et 148.)

Cédric Messiant. A Subcategorization Acquisition System for French Verbs. Dans *Proceedings of the ACL-08 : HLT Student Research Workshop*, pages 55–60, Columbus, Ohio, June 2008. Association for Computational Linguistics. (Cité pages 82 et 147.)

Cédric Messiant, Kata Gábor, et Thierry Poibeau. Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français. *Traitement Automatique des Langues*, 2010. (Cité pages 117 et 120.)

Cédric Messiant, Takuya Nakamura, et Stavroula Voyatzi. La complémentarité des approches manuelle et automatique en acquisition lexicale. Dans *Actes de la 16ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Senlis, 2009. (Cité pages 26 et 27.)

George A. Miller. Wordnet : a lexical database for english. *Communications of the ACM*, 1995. (Cité pages 16, 114 et 144.)

Philip Miller. Compléments et circonstants : distinction syntaxique ou sémantique ? Dans J.-C.Souesme, éditeur, *Actes de l'Atelier de Linguistique. SAES 1997. Cycnos*, volume 15, Nice, 1997. (Cité page 11.)

Erwann Moreau, Isabelle Tellier, Antonio Balvet, Grégoire Laurence, Antoine Rozenknop, et Thierry Poibeau. Annotation fonctionnelle de corpus arborés avec des champs aléatoires conditionnels. Dans *Actes de la 16ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Senlis, 2009. (Cité page 2.)

Ruth O'Donovan, Michael Burke, Aoife Cahill, Josef van Genabith, et Andy Way. Large-scale induction and evaluation of lexical resources from the penn-ii and penn-iii treebanks. *Computational Linguistics*, 31(3) :329–366, 2005. (Cité page 37.)

Diarmuid Ó Séaghdha et Ann Copestake. Semantic Classification with Distributional Kernels. Dans *Proceedings of COLING*, pages 649–656, 2008. (Cité pages 115, 144, 147, 150 et 154.)

Martha Palmer, Daniel Gildea, et Paul Kingsbury. The proposition bank : An annotated corpus of semantic roles. *Computational Linguistics*, 3(1) :71–106, 2005. (Cité pages 114 et 144.)

Thierry Poibeau et Cédric Messiant. Do We Still Need Gold Standard for Evaluation ? Dans *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marrakech, 2008. (Cité page 45.)

Carl Pollard et Ivan A. Sag. Information-based syntax and semantics. *Fundamentals. CLSI Lecture Notes 13*, 1, 1987. (Cité page 11.)

Judita Preiss, Ted Briscoe, et Anna Korhonen. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. Dans *Proceedings of the Meeting of the Association for Computational Linguistics*, pages 912–918, Prague, 2007. (Cité pages 51, 54, 74, 95 et 107.)

James Pustejovsky. *The generative lexicon*. The MIT Press, Cambridge, 1995. (Cité page 1.)

Philip Resnik. Selectional preference and sense disambiguation. Dans *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics : Why, What, and How ?*, pages 52–57, Washington, DC, 1997. (Cité page 124.)

Benoît Sagot. *Analyse automatique du français : lexiques, formalismes, analyseurs*. PhD thesis, Université Paris VII, 2006. (Cité page 34.)

Benoît Sagot. The lefff, a freely available, accurate and large-coverage lexicon for french. Dans *Proceedings of LREC 2010*, La Valette, Malte, 2010. (Cité page 33.)

Benoît Sagot, Lionel Clément, Eric de La Clergerie, et Pierre Boullier. The lefff 2 syntactic lexicon for French : architecture, acquisition, use. Dans *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Genua (Italy), 2006. (Cité page 33.)

Benoît Sagot et Laurence Danlos. Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire. Dans *Actes du colloque DLTAF 2006 (Description Linguistique pour le Traitement Automatique du Français) du congrès de l'ACFAS*, Montréal, Canada, 2006. (Cité page 34.)

Benoît Sagot et Eric de La Clergerie. Error mining in parsing results. Dans *Proceedings of ACL-CoLing 06*, pages 329–336, Sydney, Australie, 2006. (Cité page 34.)

Benoît Sagot et Elsa Tolone. Intégrer les tables du lexique-grammaire à un analyseur syntaxique robuste à grande échelle. Dans *Actes de la 16ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Senlis, France, 2009. (Cité pages 29, 34 et 103.)

Benoît Sagot et Darja Fiser. Construction d'un wordnet libre du français à partir de ressources multilingues. Dans *Actes de la 15ème Conférence sur le traitement automatique des langues naturelles (TALN)*, Avignon, France, 2008. (Cité page 125.)

Patrick Saint-Dizier. Verb semantic classes based on 'alternations' and wordnet-like criteria. Dans Patrick Saint-Dizier, éditeur, *Predicative Forms in Natural language and lexical Knowledge Bases*, pages 247–279. Kluwer Academic, 1998. (Cité page 145.)

Morris Salkoff. *Analyse syntaxique du français. Grammaire en chaîne.* John Benjamins, Amsterdam, 1979. (Cité page 36.)

Morris Salkoff et André Valli. A dictionary of french verbal complementation. Dans *Actes de Language and Technology Conference. Human Language and Technologies as a Challenge for Computer Science and Linguistics. In memory of M. Gross and A. Zampolli*, Poznan, Poland, 2005. (Cité page 36.)

Anoop Sarkar et Daniel Zeman. Automatic extraction of subcategorization frames for czech. Dans *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, pages 691–697, Saarbrücken, Germany, 2000. (Cité page 37.)

Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. Dans *International Conference on New Methods in Language Processing*, Manchester, UK, 1994. (Cité page 59.)

Sabine Schulte im Walde. Clustering verbs semantically according to their alternation behaviour. Dans *Proceedings of COLING*, Saarbrücken, 2000. (Cité pages 3, 115 et 116.)

Sabine Schulte im Walde. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. Dans *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas de Gran Canaria, Spain, 2002. (Cité page 54.)

Sabine Schulte im Walde. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2) :159–194, 2006. (Cité pages 115, 144, 147 et 154.)

Sabine Schulte im Walde. The induction of verb frames and verb classes from corpora. Dans *Corpus Linguistics. An International Handbook*, Berlin, 2009. Mouton de Gruyter. (Cité pages 54 et 141.)

Sabine Schulte im Walde et Chris Brew. Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information. Dans *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 223–230, Philadelphia, PA, 2002. (Cité page 116.)

András Serény, Eszter Simon, et Anna Babarczy. Automatic acquisition of hungarian subcategorization frames. Dans *Proceedings of the 9th International Symposium of Hungarian Researchers on Computational Intelligence*, 2008. (Cité page 54.)

Lei Shi et Rada Mihalcea. Putting Pieces Together : Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. Dans *Proceedings of CICLing*, pages 100–111, 2005. (Cité pages 115 et 144.)

Harold Somers. On the validity of the complement-adjunct distinction in valency grammar. *Linguistics*, 22 :501–530, 1984. (Cité page 11.)

Karen Sparck-Jones et Julia R. Galliers. *Evaluating Natural Language Processing Systems : An Analysis and Review*. Springer, Oxford University Press, Berlin, 1996. (Cité page 43.)

Suzanne Stevenson et Xavier Carreras, éditeurs. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*. ACL, Boulder, Colorado, 2009. (Cité page 2.)

Lin Sun et Anna Korhonen. Improving Verb Clustering with Automatically Acquired Selectional Preferences. Dans *Proceedings of EMNLP*, pages 638–647, 2009. (Cité pages 115, 144, 145, 147, 148, 149, 150, 151, 152 et 154.)

Lin Sun, Anna Korhonen, et Yuval Krymolowski. Verb class discovery from rich syntactic data. *LNCS*, 4919 :16, 2008. (Cité pages 145, 148, 151 et 153.)

Lin Sun, Thierry Poibeau, Anna Korhonen, et Cédric Messiant. Investigating the Cross-Linguistic Potential of VerbNet-Style Classification. Dans *Proceedings of CoLing*, 2010. (Cité pages 117 et 120.)

Mihai Surdeanu, Sanda M. Harabagiu, John Williams, et Paul Aarseth. Using predicate-argument structures for information extraction. Dans *ACL*, pages 8–15, 2003. (Cité pages 7, 23 et 125.)

Yoshimi Suzuki et Fumiyo Fukumoto. Classifying Japanese Polysemous Verbs based on Fuzzy C-means Clustering. Dans *Proceedings of TextGraphs-4*, pages 32–40, 2009. (Cité pages 115, 144 et 154.)

Robert Swier et Suzanne Stevenson. Unsupervised Semantic Role Labelling. Dans *Proceedings of EMNLP*, 2004. (Cité pages 115 et 144.)

Lucien Tesnière. *Éléments de Syntaxe structurale*. Klincksieck, Paris, 1959. (Cité pages 9 et 62.)

Akira Ushioda, David A. Evans, Ted Gibson, et Alex Waibel. The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. Dans *Proceedings of the SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text*, pages 95–106, Columbus, Ohio, 1993. (Cité page 50.)

Karel van den Eynde et Claire Blanche-Benveniste. Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale. *Cahiers de Lexicologie*, 32 :3–27, 1978. (Cité page 30.)

Karel van den Eynde et Piet Mertens. La valence : l'approche pronominale et son application au lexique verbal. *French Language Studies*, 13(1) :63–104, 2003. (Cité pages 30 et 31.)

Karel van den Eynde et Piet Mertens. *Le dictionnaire de valence Dicovalence : manuel d'utilisation*. Manuscript, Leuven, 2006. (Cité pages 30 et 95.)

Gloria Vázquez, Ana Fernández, Irene Castellón, et M. Antonia Martí. Clasificación verbal : Alternancias de diátesis. Dans *Quaderns de Sintagma*. Universitat de Lleida, 2000. (Cité page 145.)

Deepak Verma et Marina Meila. A comparison of spectral clustering algorithms. Rapport technique, Department of CSE University of Washington Seattle, 2005. (Cité page 149.)

Andreas Vlachos, Anna Korhonen, et Zoubin Ghahramani. Unsupervised and constrained dirichlet process mixture models for verb clustering. Dans *Proceedings of the Workshop on on GEMS*, pages 74–82, 2009. (Cité pages 115, 144, 147 et 154.)

Ulrike von Luxburg. A tutorial on spectral clustering. *STAT COMPUT*, 17 : 395 – 416, 2007. (Cité page 149.)

Piek Vossen. *EuroWordNet : A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998. (Cité page 145.)

Wendy Wilkins, éditeur. *Thematic relations*, volume 21. Academic Press, 1988. (Cité page 13.)

Beñat Zapirain, Eneko Agirre, et Lluís Màrquez. Robustness and Generalization of Role Sets : PropBank vs. VerbNet. Dans *Proceedings of ACL*, 2008. (Cité pages 115 et 144.)

Lihi Zelnik-Manor et Pietro Perona. Self-tuning spectral clustering. *NIPS*, 17 (1601-1608) :16, 2004. (Cité pages 117, 148 et 149.)

Ce document a été préparé à l'aide des éditeurs de texte Geany et Kile et du logiciel de composition typographique LaTeX 2ε.

Titre Acquisition automatique de schémas de sous-catégorisation à partir de corpus bruts

Résumé Cette thèse de doctorat traite de l'acquisition automatique d'informations lexicales à partir de corpus. Nous nous intéressons en particulier à l'acquisition de schémas de sous-catégorisation de verbes pour le français. Nous avons mis au point un système permettant d'acquérir automatiquement ce type d'informations. L'utilisation de ce système sur un corpus de très grande taille a permis d'acquérir un lexique de sous-catégorisation pour le français. L'évaluation de cette ressource et sa comparaison à d'autres ressources montre l'apport des méthodes automatiques dans la constitution de lexiques. Nous montrons par ailleurs l'adaptabilité de tels systèmes et leur aptitude à acquérir des ressources « spécialisées » rapidement. Nous explorons enfin l'utilisation de telles ressources pour l'acquisition de classes sémantiques de verbes dans la lignée des travaux de Beth Levin, ce qui permet d'étudier les liens entre syntaxe et sémantique.

Mots-clés verbe, sous-catégorisation, lexique, acquisition automatique, syntaxe, classes sémantiques

Title Automatic acquisition of subcategorization frames from raw corpora

Abstract This thesis deals with the automatic acquisition of lexical information from corpora. We are especially interested in the acquisition of subcategorization frames for French verbs. We designed and developed a system to acquire automatically such information. Using this system on a very large corpus, we produced a subcategorization lexicon for French. The evaluation of this resource and its comparison with other resources shows the interest of of automatic methods for the creation of lexicons. We show that the adaptability of such systems and their ability to acquire quickly "specialized" linguistic resources. Finally, we explore the use of such resources for the acquisition of Levin style semantic verb classes, therefore we study the relationship between syntax and semantics.

Keywords verb, subcategorization, lexicon, automatic acquisition, syntax, semantic classes